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

Volume 17 Issue 4
April 2026



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

THE SCIENCE AND INFORMATION ORGANIZATION

www.ijacsa.thesai.org

Table of Contents

Paper 1: EU Public Procurement Contract Value Prediction: A Machine Learning Approach

Authors: Muhammad Azizur Rahman, Anna Hlinska

PAGES 1 - 7

Paper 2: Unified Modular Architectural and Software Model for Educational XR Systems

Authors: Miroslav GALABOV

PAGES 8 - 14

Paper 3: Interruptible Multi-Agent Debate: Sentence-Level Disclosure and Urgency-Based Turn-Taking for Early Error Correction

Authors: Akikazu Kimura, Ken Fukuda, Yasuyuki Tahara, Yuichi Sei

PAGES 15 - 25

Paper 4: Secure User Authentication Model Using Identity-Based Encryption (IBE) Scheme: Challenges, Techniques, and Trends

Authors: Raja Farah Sharima Raja Muhamad Danial, Nazhatul Hafizah Kamarudin, Abdul Ghafar Jaafar

PAGES 26 - 36

Paper 5: Improving Heart Sound Diagnosis with a Combined CNN-LSTM and Dual-Attention Deep Learning Model

Authors: Arshad Jamal, R. Kanesaraj Ramasamy, Junaidi Abdullah

PAGES 37 - 46

Paper 6: Application of Large-Scale Array Acoustic System Performance Detection Technology Based on Embedded ARM7 + uClinux Platform in Petroleum Exploration

Authors: Zhiyuan Sun, Cheng Yang, Guixu Xu

PAGES 47 - 57

Paper 7: Factors Influencing Generative AI-Enabled e-Government Services (GAIGS) Information Quality: A Systematic Literature Review

Authors: Azwan Abd Aziz, Rozi Nor Haizan Nor, Yusmadi Yah Jusoh, Wan Nurhayati Wan Ab. Rahman, Khairi Azhar Aziz, Nur Ilyana Ismarau Tajuddin, Raditya Muhammad

PAGES 58 - 69

Paper 8: Balancing Accuracy Robustness and Explainability in E-Commerce Recommender Systems

Authors: Mansor Alohali

PAGES 70 - 81

Paper 9: An Algorithm for Image Dataset Compression and Privacy Enhancement via Fusing Bilateral Filtering and Easy-to-Complex Trajectory Matching Distillation

Authors: Qiao Zhou, Lei Zhang, Longjie Li, Tong Wang, Jun Cheng

PAGES 82 - 89

Paper 10: Cryptographic Vulnerability Assessment and Secure Redesign of an ECC-Based Authentication Framework for Energy Internet Vehicle-to-Grid Systems

Authors: Haewon Byeon

PAGES 90 - 97

Paper 11: An AI-Powered Approach for Medical Specialty Triage Using Natural Language Processing and Transformer Models

Authors: Anas Chahid, Ismail Chahid, Wafae Mrabti, Mohamed Emharraf, Mohammed Ghaouth Belkasmi

PAGES 98 - 107

Paper 12: Multi-View Behavioral Probing for Political Bias in Arabic and Multilingual Transformers Before and After Domain Adaptation

Authors: Ahmad Abdelhameed, Ensaf Hussein Mohamed, Walaa Medhat

PAGES 108 - 118

Paper 13: Trust and Hallucinations: A Study of 39 Experts on AI-Assisted Requirements Reverse Engineering

Authors: Abdullah A H Alzahrani

PAGES 119 - 128

Paper 14: Generative Monocular Perception Pipeline-Based Framework for Accurate Stem Detection in Automated Strawberry Harvest

Authors: Kohei Arai, Jin Sawada, Mariko Oda

PAGES 129 - 139

Paper 15: Method for Improving Object Detection and Classification Accuracy Using a Small Training Dataset by Reducing the Number of Classes

Authors: Kohei Arai

PAGES 140 - 146

Paper 16: Enhancing Communication Accessibility: Real-Time Recognition and Synthesis of Arabic Sign Language Gestures Using Long Short-Term Memory

Authors: Mina Nagy Gaber Sorial, Rodaina Abdelsalam, Mayar Ali, Hesham Hassan

PAGES 147 - 154

Paper 17: A Multi-Layer Computational Framework for Predicting Student Performance Ranges in Higher Education Using Machine Learning

Authors: Abdellatif HARIF, Moulay Abdellah KASSIMI

PAGES 155 - 164

Paper 18: Digital Infrastructure Transformation in the Public Sector: Explaining IPv6 Adoption Through the UTAUT Framework

Authors: Aydın Koçak, Uğur Dagtekin, Ahmet Kamil Kabakuş

PAGES 165 - 176

Paper 19: FIFO Age-Cohort Stochastic MILP for Perishable Inventory Optimization

Authors: Hirman Rachman, Saib Suwilo, Sutarman, Elvina Herawati

PAGES 177 - 187

Paper 20: Detection of Video Anomalies via CNN-LSTM Model for Intelligent Surveillance

Authors: Mohamed H. Mousa, Yasser M. Ayid, Ayman E. Khedr, Ahmed M. Elshewey

PAGES 188 - 197

Paper 21: Evidence-Driven AI Governance for Healthcare: A PEARL-PATHWAY Analysis of Madinah

Authors: Roba Alsaigh, Rashid Mehmood, Iyad Katib, Abdulaziz A. Almuzaini,
Sami Saad Albouq, Sami Alshmrany

PAGES 198 - 214

Paper 22: KnowRAG: A Zero-Shot Diagnostic Analysis of Knowledge Base Coverage in Scientific Retrieval-Augmented Generation

Authors: Assmaa MOUTAOUKKIL, Ali EL MEZOUARY, Kaoutar BOUMALEK

PAGES 215 - 224

Paper 23: Determinants of Telemedicine Security Readiness Among ICT Professionals: Extended UTAUT Model

Authors: Fazlina Mohd Ali, Afraa Mohammed Mohammed Gaashan,
Marizuana Mat Daud, Rimaniza Zainal Abidin

PAGES 225 - 230

Paper 24: Evaluating Perceptual Reliability of Latent Attribute Control in Diffusion-Based Fashion Generation

Authors: Noriaki Kuwahara, Shintaro Kawanami, Takashi Sato, Dongeun Choi

PAGES 231 - 240

Paper 25: Performance Comparison of Regularization Methods on Transfer Learning Algorithm for Fish Species Classification

Authors: Handrie Noprisson, Anita Ratnasari, Sri Dianing Asri, Vina Ayumi, Hadiguna Setiawan

PAGES 241 - 247

Paper 26: FEM-KP: A Functional Evaluation Metric for Keyphrase Prediction Models

Authors: Lahbib Ajallouda, Ahmed Zellou

PAGES 248 - 257

Paper 27: A Neutrosophic Machine Learning-Based Intelligent Ensemble Model for Sustainable Tea Yield Prediction Under Climatic Variability

Authors: Maitraya Dey, Pushpita Roy, Shubhendu Banerjee, Amrut Ranjan Jena, Rakesh Naskar,
Suparna Dasgupta, Soumyabrata Saha, Sudarshan Nath, Bikash Mondal

PAGES 258 - 265

Paper 28: Incorporating Generative AI in Foreign Language Teaching Preparation: An Empirical Investigation Into the Effects of AI-Assisted Lesson Planning

Authors: Xinxin Guo, Hao Chen

PAGES 266 - 275

Paper 29: Climate Change on Social Media: AI and Deep Learning-Based Analysis of Tweets

Authors: Bahar URHAN, Mehmet KAYAKUŞ, Dilşad ERDOĞAN, Gülten ADALI, Emrah BOZKURT, Zeynep Nihan BAKIR

PAGES 276 - 287

Paper 30: Preference-Controllable Multi-Objective Deep Reinforcement Learning for Human-Robot Task Allocation in Service Environments

Authors: Asmaa Rashed Alahmari, Wadee Alhalabi

PAGES 288 - 299

Paper 31: Fuzzy-Integrated Modular Neural Networks for Accurate Prediction of On-Time-In-Full Supply Chain Performance

Authors: Mariem Mrad, Mohamed Amine Frikha, Younes Boujelbene, Soufiene Ben Othman

PAGES 300 - 308

Paper 32: Federated and Secure Artificial Intelligence as a Driver of Operational Excellence and Competitive Advantage in Smart Manufacturing Enterprises

Authors: Adel Saad Assiri

PAGES 309 - 317

Paper 33: Blockchain Governance Framework and Assessment Tools from a Readiness Perspective

Authors: Nur Indrawati, Dana Indra Sensuse, Deden Sumirat Hidayat, Erisvaha Kiki Purwaningsih

PAGES 318 - 330

Paper 34: The Untapped Potential of Extended Reality for Indigenous Medicinal Knowledge: A Review of Cross-Disciplinary XR Applications

Authors: Giampearo Anak Peter, Suriati Khartini Jali, Jane Labadin, Abby Lian Hendrick, Ally Dian Hendrick, Nurul Farizah Ridzuan

PAGES 331 - 338

Paper 35: IFDA-EMA-YOLOv9: Wheat Disease Detection Integrating Flow Optimization and Auxiliary Supervision

Authors: Jiaxiang Fan, Leixiao Li

PAGES 339 - 352

Paper 36: Energy-Efficient Multi-Hop LoRa Communication in Forested Environments via Proximal Policy Optimization

Authors: Muhd Kahfi bin Jumali, Lim Kit Guan, Ervin Gubin Moung, Lorita Angeline, Tianlei Wang, Kenneth Teo Tze Kin

PAGES 353 - 363

Paper 37: Comparative Analysis of Fixed vs Machine Learning Dynamic Pricing Models: A Computational Performance Study

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

PAGES 364 - 370

Paper 38: Machine Learning-Driven Resource Provisioning in Modern Cloud Environments: A Taxonomic Survey

Authors: Stefanus Albert Kosim, Bagus Jati Santoso, Deka Julian Arrizki, Riki Mi'roj Achmad, I Nyoman Gede Artadana Mahaputra Wardhiana, Royyana Muslim Ijtihadie

PAGES 371 - 383

Paper 39: A Machine Learning-Driven Framework for Accurate Brain Image Registration in Multimodal and Noisy Environments

Authors: M. S. Minu, Mutharasu M, S. Hemamalini, Sunitha T, Mohanaprakash T A, Justindhas Y

PAGES 384 - 394

Paper 40: On the Security of Authentication Protocols for Remote Healthcare Systems Through Cryptographic Vulnerability Analysis and Secure Protocol Redesign

Authors: Haewon Byeon

PAGES 395 - 402

Paper 41: An Ensemble Boosting Approach with Boruta Feature Selection for Predicting E-Payment Adoption

Authors: Mariana Purba, Junaidi Junaidi, Lemi Iryani, Nia Umilizah, Handrie Noprisson, Nur Ani

PAGES 403 - 409

Paper 42: Beyond Paper and Pencil: Evaluating the Effectiveness and User Perception of a Digital Reading Proficiency Assessment Platform

Authors: Ivy M. Tarun

PAGES 410 - 417

Paper 43: Hybrid Learning-to-Rank Approach for Complex Information Retrieval Systems

Authors: Fatma Zohra Bessai-Mechmache, Yasmine Hanifi, Damia Lyna Ait Idir

PAGES 418 - 425

Paper 44: Optimizing Rainfall Prediction in Settat, Morocco, Through Machine Learning

Authors: Oussama Zemnazi, Sanaa El Filali, Sara Ouahabi, Abderrahim Mouhtadi

PAGES 426 - 433

Paper 45: A Systematic Literature Review on Artificial Intelligence Applications for Breast Cancer Classification

Authors: Nursakinah Abdullah, Qi Wei Oung, Chee Chin Lim, Chiew Chea Lau, Vrshni Menaka R Siva Nathan, Hui Wen Tiu

PAGES 434 - 448

Paper 46: Integrating Big Data and Machine Learning for Effective Cyberattack Prediction in e-Health Information Systems

Authors: Mohamed Abdelbaki, Latif Adnane, Charaf Eddine Ait Zaouiat

PAGES 449 - 454

Paper 47: Solar Irradiance Forecasting Approaches Based on Machine Learning: A Systematic Literature Review

Authors: Sempe Thom Leholo, Chunling Tu, Topside Ehleketani Mathonsi

PAGES 455 - 475

Paper 48: Efficient Computation of Parametric Exponentiation in Parametric Algebra

Authors: Khudoykulov Zarifjon, Khudoynazarov Umidjon, Muminova Mastura

PAGES 476 - 487

Paper 49: An Interoperable Multi-Agent Architecture for Personalized Smart Learning Using Generative AI and Learning Analytics

Authors: Al Mahdi Khaddar, Youssef Said, Amine Dehbi, Tarik Chafiq

PAGES 488 - 500

Paper 50: Mapping Tourist Sentiments Through Lexicon-Based Analysis of Social Media Reviews: The Case of Salak Sibetan Agritourism

Authors: Paula Dewanti, Putu Adi Guna Permana, I Gusti Ayu Widari Upadani,
Ellyn Ly Maramento, Alfred John G. Borreros

PAGES 501 - 506

Paper 51: Optimization of Access Point Location Using an Integer Programming Model

Authors: Joan Marie M. Panes, Marilou O. Espina, Jovelin M. Lapates

PAGES 507 - 513

Paper 52: AI and Blockchain for Secure Healthcare Data Management: A Bibliometric Analysis of Research Trends and Thematic Clusters (2020–2025)

Authors: Maroufi Mohammed, Lamzabi Siham, Ziti Soumia

PAGES 514 - 522

Paper 53: A Bidirectional LSTM–Sentiment Fusion Framework for Dynamic Financial Market Prediction

Authors: Minal Dhankar, Neha Gupta

PAGES 523 - 533

Paper 54: A Systematic Review of Graph Neural Networks and Social Network Analysis Techniques for Public Sentiment Uncovering

Authors: Adi Wibowo, Wijayanto, Henri Tantyoko, Ari Wibisono, Usman Ependi

PAGES 534 - 545

Paper 55: AI-Based Process Mining Framework for Process Business Integration in an Enterprise System

Authors: Mardhani Riasetiawan, Ahmad Ashari

PAGES 546 - 552

Paper 56: Intelligent ECU Load Management in Electric Vehicles Using a Gated Multi-Stage Machine Learning Framework

Authors: Vaishali Mishra, Sonali Kadam

PAGES 553 - 569

Paper 57: Hybrid Feature Learning with TF-IDF and SBERT for Ambiguous Requirement Classification

Authors: Fariha Khalid, Muhammad Yaseen, Gohar Rahman, Nauman Mazhar,
Muhammad Asif Nauman, Aida Mustapha

PAGES 570 - 576

Paper 58: QR Code-Based Access Control Systems: Architectural Taxonomy, Security Landscape, and Future Research Directions

Authors: Worood Alsawi, Dina M. Ibrahim

PAGES 577 - 592

Paper 59: Real-Traffic-Trained Intelligent IDS for Advanced Cyberattack Detection in Enterprise Networks

Authors: Dalila Naira Chinchay, Rodrigo Calderón Ari, Liset S. Rodriguez-Baca

PAGES 593 - 602

Paper 60: Evaluating Open-Source LLMs for Thai Clinical Information Extraction

Authors: Somkiat Kosolsombat, Phatnattachat Chatsiraphon, Taratep Si-Aksorn, Chiabwoot Ratanavilisagul

PAGES 603 - 608

Paper 61: Two-Phase Transfer Learning Framework for Automated Depression Classification in the Elderly via Facial Expression Recognition

Authors: Muhammad Daffa Zahrandika Wibisono, Marizuana Mat Daud, Wan Mimi Diyana Wan Zaki

PAGES 609 - 617

Paper 62: Comparison of Time-Domain and Frequency-Domain EMG Features for Gait Phases Classification Using Machine Learning

Authors: Muhamad Amirul Sunni Rohim, Nurhazimah Nazmi, Shin-Ichirou Yamamoto, Muhammad Kashfi Shabdin, Mohd Asyadi Azam

PAGES 618 - 626

Paper 63: A Framework for Digital Technology and AI Adoption in Slovak Firms: Evidence from Qualitative Analysis

Authors: Martina Chrancokova, Ludmila Mitkova, Lukas Vartiak

PAGES 627 - 636

Paper 64: Feature-Level Analysis and Robust Baselines for EEG-Based Imagined Speech Recognition on the ASU Dataset

Authors: Hatem T M Duhair, Masrullizam Mat Ibrahim, Jamil Abedalrahim Jamil Alsayaydeh, Mazen Farid, Safarudin Gazali Herawan

PAGES 637 - 651

Paper 65: A Multivocal Literature Review of Metadata Governance: Conceptual Foundations and Research Gaps

Authors: Dana Indra Sensuse, Alivia Yulfitri, Erisva Hakiki Purwaningsih, Anton Satria Prabuwono

PAGES 652 - 663

Paper 66: Advanced Forensic Analysis Techniques for Insider Threat Detection in Database Management Systems

Authors: Kholod Saeed Talea AlQahtani, Mounir Frikha, M. M. Hafizur Rahman

PAGES 664 - 671

Paper 67: A Hybrid Explainable Ensemble Learning Framework for Health Risk Prediction

Authors: Hussain AlSalman

PAGES 672 - 681

Paper 68: A Digital Twin-Enabled Approach to Optimize Freight Fleet Operations in a Peruvian Transportation Company

Authors: Deyber Flores Cabezas, Liset S. Rodriguez-Baca

PAGES 682 - 689

Paper 69: Real-Time Thought-to-Vision Generation Using Low-Channel EEG and Feature-Fusion Learning

Authors: Abha Marathe, Medha Wyawahare, Milind Rane, Vrinda Parkhi, Milind Kamble, Anvita Barlingay, Anushka Goswami, Aditya Awate

PAGES 690 - 701

Paper 70: Adoption of Blockchain Technology in Electronic Records Management in the Malaysian Public Sector

Authors: Aida Ruzana Ahmad Yani, Umi Asma' Mokhtar

PAGES 702 - 714

Paper 71: Reproducible Prediction Framework of Customer Churn Using Machine Learning, Advanced Data Science and Business Intelligence Techniques

Authors: Younes KOULOU, Norelislam EL HAMI

PAGES 715 - 722

Paper 72: Attribute-Conditioned Attention Scaling for Text-to-Image Diffusion Models

Authors: Rabia Tahir, Bilal Ahmed Memon

PAGES 723 - 729

Paper 73: Adaptive Multi-Layer Encryption for Enhancing Security in Smart Home IoT Ecosystems

Authors: Dancan Obuya Machuki, Kennedy Ronoh

PAGES 730 - 736

Paper 74: Workload-Aware Storage Reduction for Multi-Tenant SIEM on ClickHouse

Authors: Nutthakorn Chalaemwongwan

PAGES 737 - 745

Paper 75: Web Server Seizure and Live Acquisition: A Legally Compliant Forensic Framework for Indonesia

Authors: Irwan Hariyanto, Yudi Prayudi, Rimba Whidiana Ciptasari

PAGES 746 - 757

Paper 76: Adaptive Phishing Website Detection Using Incremental Machine Learning: A Dynamic Approach to Cybersecurity Threats

Authors: Ajla Kulaglic, Mutaz A. B. Al-Tarawneh

PAGES 758 - 773

Paper 77: Automated Interactance Near-Infrared Spectral Acquisition System for Mandarin Quality Assessment

Authors: Van-Linh Lam, Dinh-Tri Nguyen, Thanh-Trung Le, Hoang-Tien Nguyen, Phuoc-Loc Nguyen, Quoc-Khanh Huynh, Nhut-Thanh Tran, Chanh-Nghiem Nguyen

PAGES 774 - 783

Paper 78: The Impact of Modern AI on Software Development: A Systematic Literature Review

Authors: Meryem Bensaid, Marouane Achbari, Younesse Ouahbi, Soumia Ziti

PAGES 784 - 791

Paper 79: A Survey on Extended Reality Technologies for Industrial Maintenance

Authors: Mouad Danane, Abdelmajid El Ouadi, Youssef Rochdi

PAGES 792 - 816

Paper 80: Design and Implementation of a Trust-Aware Fog Computing Simulation Framework Using WorkflowSim

Authors: Chia Chuan Wu, Selvakumar Manickam, Shams Ul Arfeen Laghari, Shankar Karuppayah

PAGES 817 - 825

Paper 81: User-Centered Ergonomic Design of a Portable EEG System for Evaluating Dolphin-Assisted Therapy in Children with Neurodevelopmental Disorders

Authors: Brenda Lorena Flores Hidalgo, Urim Rarael Pérez Bernal, Jesús Jaime Moreno Escobar, Hugo Quintana Espinosa, Ana Lilia Coria Páez

PAGES 826 - 836

Paper 82: Automated Medical Image De-Identification via U-Net++ Segmentation and Conditional GAN Inpainting

Authors: Ismail Chahid, Anas Chahid, Yassine Chahid, Aissa Kerkour Elmiad, Mohammed Badaoui

PAGES 837 - 846

Paper 83: Learning Structural Regularities over Mobile UI Flows from Interaction Traces

Authors: Iskandar Salama, Masayasu Atsumi

PAGES 847 - 865

Paper 84: Risk Assessment and Risk Management Challenges in Intelligent IoT-Based Smart City Infrastructures

Authors: Abdullah Alessa, Yaseen Alduwayl, M M Hafizur Rahman

PAGES 866 - 878

Paper 85: A Cryptographic Framework Using AES-ECC with Threshold Key Management for Cloud Storage Systems

Authors: Abdulsalam Ibrahim Almirdasi, Mohamed Tahar Ben Othman

PAGES 879 - 887

Paper 86: AFT-Attentive BiLSTM: Improving Early Warning of Firm Financial Distress with Temporal Attention in an Accelerated Failure Time Framework

Authors: Muhammad Ali Chohan, Suresh Ramakrishnan, Mohammad Abrar, Shamaila Butt, Shahid Kamal

PAGES 888 - 899

Paper 87: TrustPatch-X: Multi-Stage Explainable Framework for Reliable LLM Patch Validation

Authors: Sheetal Madhukar Parate, Jasmine Selvakumari Jeya I

PAGES 900 - 908

Paper 88: An Identity-Aware Privacy-Preserving Deep Learning Framework for Culturally Sensitive Image Sharing

Authors: Mahmoud Obaid, Hadeel Bkhaitan, Duha Maali, Saja Hammad, Thaer Thaher

PAGES 909 - 919

Paper 89: Stability-Weighted Feature Selection with Adaptive PSO-SGD for Neural Network-Based Predictive Hiring

Authors: Yassine Temsamani Khallouk, Said Achchab

PAGES 920 - 933

Paper 90: GTMedoids: A New Grey Sheep Users Detection Approach

Authors: Bouchra Boualaoui, Ahmed Zellou, Lahbib Ajallouda

PAGES 934 - 942

Paper 91: A Proxy-Based DevSecOps Framework for Multi-Tier Web Applications

Authors: Abderrahim Rida, Abdelaziz Bakhil, Ayoub Ait Lahcen

PAGES 943 - 955

Paper 92: Hybrid Shape Descriptor Fusion with LightGBM for Robust 3D Mesh Classification and Retrieval

Authors: Khadija Arhid, Youness Ghazi, Ilham Kachbal, Fatima Rafii Zakani, Mohcine Bouksim, Said El Abdellaoui, Taoufiq Gadi, Mohamed Aboulfatah

PAGES 956 - 966

Paper 93: A Survey of FPGA Floorplanning for Dynamic Partial Reconfiguration: From Heuristic Approaches to Autonomous AI-Driven Methods

Authors: Ibrahim LIMEM, Sadok BAZINE, Abdessalem BEN ABDELALI

PAGES 967 - 975

Paper 94: SentimentPulse: A Concurrent Multi-Platform System for Social Media Sentiment Monitoring with LLM-Based Interpretation

Authors: Gabriel A. León-Paredes, Erika C. Villa-Quishpi, Jorge E. Márquez-Chávez, Erick F. Zhigue-Granda

PAGES 976 - 984

Paper 95: Improving Performance and Accuracy in Decision Trees: A Literature Survey on Impurity Functions

Authors: Abed Alsulami, Reda Khalifa, Wajdi Alghamdi

PAGES 985 - 1000

EU Public Procurement Contract Value Prediction: A Machine Learning Approach

Dr Muhammad Azizur Rahman, Anna Hlinska
Department of Computer Science, Cardiff Metropolitan University, Llandaff Campus, Western Avenue, Cardiff, CF5 2YB, UK

***Abstract*—Using modern analytical tools, such as machine learning, organisations can collect and analyse large amounts of data related to suppliers, pricing, demand structure, and market trends. This study uses machine learning models to predict the value of EU public procurement contracts based on the Tenders Electronic Daily (TED) database. The analysis covers 13,345,120 initial contract records for the period 2006-2025 in 33 states and 63 procurement sectors. After careful data quality control procedures, the analytical set was formed by 10,038,018 valid contracts (75% retention rate). Three complementary methodologies were used: Random Forest regression to identify nonlinear patterns, ordinary least squares (OLS) regression to interpret coefficients, and K-means clustering to classify procurement behaviour at the country level. The Random Forest achieved cross-validation R^2=0.2795 and the test R^2=0.2613, with the country of origin dominating the predictive value (Germany: 24.39%, United Kingdom: 2.33%, Italy: 1.96%). Temporal features had a significance of 18.36%, while competition indicators (number of proposals: 8.99%) and structural characteristics (batch number: 12.29%) had a significant impact. The OLS regression showed statistically significant effects for countries: Germany showed 98.4% lower contract prices despite being Europe's largest economy, reflecting federal administrative fragmentation, while Italy showed 295.7% higher rates thanks to centralised infrastructure projects. The K-means clustering revealed three clear procurement profiles: Greece as a transparency-focused outsider (109 average bids per contract), 19 mature economies with high value, and 13 less valuable, fragmented systems, including Germany. The results show that institutional frameworks dominate economic factors in determining the value of contracts, which has political implications for the design of procurement systems across the EU.**

***Keywords*—Public procurement; EU procurement; machine learning; Random Forest; Contract Value Prediction; TED database; institutional economics; OLS regression; K-means clustering; predictive modelling; data quality; EU AI Act**

I. Introduction

Public procurement constitutes a significant component of public expenditure in the member states of the European Union, accounting for approximately 14% of GDP or 2 trillion euros annually [1]. The Tenders Electronic Daily (TED) database serves as the central repository for procurement information, publishing contract award notices that provide transparency into public spending decisions across 33 EU and European Economic Area countries. Despite this wealth of data, predicting contract values remains challenging due to complex interactions between country-specific institutional frameworks, sectoral characteristics, temporal trends, and competitive dynamics [2, 3]. Applying machine learning to public procurement data, budget planning, supplier strategies, and procurement policy evaluation. The scale and complexity of EU public procurement present both opportunities and challenges for data-driven analysis. Accurate contract value forecasting has significant policy and practical implications for many stakeholder groups. Procurement agencies need reliable forecasts for budget planning and resource allocation, especially against the backdrop of increased fiscal constraints in European governments. Suppliers benefit from understanding typical contract sizes when evaluating market entry decisions in different countries and sectors, which has implications for competitive strategy and capacity planning [5]. Traditional approaches to contract value analysis primarily rely on descriptive statistics and linear regression models, which may not consider the nonlinear relationships and interaction effects characteristic of procurement processes [6].

Recent advances in machine learning provide opportunities to improve prediction accuracy and uncover hidden patterns through algorithms capable of detecting complex relationships in high-dimensional datasets [7, 8]. The practical challenges of applying AI to procurement data go beyond algorithmic selection and encompass the fundamental limitations of data infrastructure. Recent comprehensive assessments show that while all EU Member States now publish structured procurement data that meets minimum machine learning standards, critical quality issues remain: approximately 30% of TED recordings contain missing values – this figure has not changed since 2017 – with supplier registration numbers missing in 86% of procedures and evaluation contracts missing in 63% of cases [22].

This study examines several critical gaps in the existing literature. First, it provides comprehensive machine learning analysis of the value of EU public procurement contracts, spanning nearly two decades (2006 – 2025) and covering all Member States, going beyond the country-specific or sectoral analyses that dominate modern research. Second, it systematically compares multiple modelling approaches – Random Forest for prediction accuracy, OLS regression for interpretive coefficients, and K-means clustering for identifying behavioural patterns — to provide both predictive capability and essential insights into the factors that determine contract value. Third, it quantifies the relative importance of country-level institutional factors compared to sectoral and temporal characteristics, which provides empirical evidence for the ongoing debate on the role of institutional quality in public procurement outcomes [12]. Fourth, it demonstrates the practical application of machine learning techniques to real-

world procurement data, which is characterised by missing values, inconsistent formats, and excessive computational loads.

The rest of this study is organised as follows: Section II reviews related work. Section III describes the research methodology. Section IV presents the results and analysis. Section V discusses the findings. Section VI outlines future directions, and Section VII concludes the study.

II. Related Work

The literature review summarises the key findings of previous studies, organised by geographical coverage, methodological approach, and content focus. Table I provides a comparative overview of key prior studies alongside this work. The latest comprehensive assessments of procurement data infrastructure show significant progress along with ongoing challenges. Longobucco and Ferverda (2025) [22] conducted a systematic assessment of the suitability of procurement data for AI-based anti-corruption analysis in all EU Member States and associated countries. Their "funnel logic" system assesses data availability at several levels: structured procurement data available en masse, beneficial ownership registers, political finance disclosures, Politically Exposed Persons (PEP) databases, procurement complaints, and media ownership records. The study confirms that all EU-27 countries are now publishing procurement data that meets minimum standards for machine learning applications, with Croatia, Estonia, and Latvia recognised as leaders in data quality and availability. However, critical gaps remain: approximately 30% of TED recordings contain missing values — this figure has not changed since 2017 — with national vendor registration numbers missing in 86% of procedures and missing contract evaluation numbers in 63% of cases [22]. These systemic data quality issues directly limit the predictive performance achieved by even complex algorithms, setting realistic expectations for model accuracy [10]. García Rodríguez et al. (2021) [11] analysed 58,337 Spanish tenders between 2012 and 2018, comparing linear regression, Random Forest, isotonic regression, and artificial neural networks to predict reward prices. Their conclusion that neural networks and isotonic regression outperformed Random Forest differs from our results, likely reflecting their focus on a single country's relatively homogeneous institutional environment, dominated by nonlinear sector-specific patterns, in contrast to the pan-European sample, where institutional heterogeneity in 33 countries contributes to Random Forest's ability to capture country-specific effects [17, 18]. Recent surveys of AI and ML applications in procurement [19], combined with country-specific studies in Thailand [20] and Portugal [21], demonstrate comparable levels of accuracy, but their narrower scope limits generalisation to the pan-European institutional diversity examined here.

This research goes beyond existing research in several dimensions. First, it provides the most comprehensive geographical and temporal coverage to date: 10 million valid contracts for the period 2006–2025 in all 33 EEA and EEA member states, compared to previous studies targeting individual countries or sectors. Second, it systematically addresses the data quality issues documented by Longobucco and Ververda (2025) [22] through explicit preprocessing procedures: deletion of 21.6% of records with missing contract values, exclusion of exceptions over €100 million, standardisation of incompatible formats, and implementation of stratified sample selection to maintain representativeness despite computational limitations. Third, it combines predictive modelling (Random Forest), causal inference (OLS regression with interpreted coefficients), and unsupervised pattern detection (K-means clustering), allowing triangulation between multiple analytical frameworks. Fourth, it quantifies the priority of institutional over economic determinants through the analysis of the importance of features, directly testing competing theoretical views on procurement management. Finally, it establishes a methodological framework aligned with the new standards for AI governance in the EU (EU Regulation 2024/1689 on the AI Law 2024/1689, MCC-AI guidelines), demonstrating how government organisations can develop sophisticated analytical capabilities using open algorithms and public datasets while maintaining transparency and vendor independence [15, 16].

TABLE I. Comparison of Contract Value Prediction Studies

Ref	Geographic Scope & Data	Approach	Algorithm(s)	Performance
[11]	Spain, 58,337 (2012–2018)	Supervised	LR, RF, Isotonic, ANN	ANN & Isotonic > RF
[19]	Multiple countries, 2024 review	Supervised	Mixed methods review	Systematic review
[20]	Thailand, 283,000 (2019)	Supervised	ANN, DT, KNN	77.8% accuracy
[21]	Portugal construction, 2015–2022	Supervised	SVM, RF, +3 others	Multiple compared
This	EU-wide 33 states, 10M (2006–2025)	Supervised + Unsupervised	RF, OLS, K-means	RF R^2=0.2613 / OLS R^2=0.1552

III. Research Methodology

The main objective of this study is to predict the value of contracts in EU public procurement using machine learning methods and to identify the institutional determinants that most significantly affect these values. The methodology is structured in four main stages: data collection and pre-processing, feature engineering, model development, and evaluation.

A total of 13,345,120 procurement contracts were downloaded and extracted from the Tenders Electronic Daily (TED) database, covering the period from 2006 to 2025 in all 33 EEC and EEA member states. TED is the official source of European public procurement reports, providing comprehensive data on contracts exceeding EU thresholds [14]. After rigorous data cleaning and validation — including the removal of records with missing contract values (21.6% of the original dataset), the elimination of exceptions exceeding €100 million, and the exclusion of incomplete records — the final dataset contained 10,038,018 valid contracts, representing a retention rate of 75%. This significant dataset provides reliable coverage of 63

procurement sectors classified under the Common Procurement Vocabulary (CPV) system, ensuring representativeness in both geographical and sectoral dimensions. These values of lack of value are closely aligned with the pan-European data quality assessments, which document approximately 30% of incomplete TED recordings, with 21-30% of cases missing contract values, and 86-89% of procedures lacking buyer information [22], confirming that problems with observed data reflect systemic infrastructure limitations rather than extraction errors.

The data extraction process used the TED API and bulk upload capabilities. The key variables that stand out are: contract value (dependent variable), country of the contracting authority, year of contract, procurement sector (CPV code), batch number within framework agreements, number of proposals received, and details of the buyer's organisation. Due to computational constraints and the need for balanced model training, a stratified random sample of 10,000 observations was taken from a cleaned dataset, which maintained proportional representation across countries, sectors, and time periods. Fig. 1 presents the methodology framework.



Fig. 1. Methodology framework.

IV. Results and Analysis

This section presents the empirical findings of our machine learning analysis on 10,000 EU public procurement contracts. We look at four key dimensions: (A) the characteristics and preparation of the dataset, (B) the importance of features and performance of the comparative visualisation model, (C) the impacts for individual countries and their institutional drivers, and (D) clustering patterns in European procurement systems.

A. Dataset Characteristics

The results are presented through tables (Table II to Table VI) and figures (Fig. 2 to Fig. 6), which together demonstrate the dominance of institutional factors over economic determinants. Table II summarises the key characteristics of the final analytical dataset.

TABLE II. Characteristics of the Dataset

Specifications	Cost
Total number of contracts	10,000
Geographical coverage	33 European countries
Time slot	2005-2023 (19 years old)
Training kit	8,000 (80%)
Test Kit	2,000 (20%)
Total number of traits (after coding)	128
Features of the country	33
Features of the sector	18
Average contract value	€9,173,061
Median contract value	€545,420

Fig. 2 shows the temporal evolution, showing steady growth from 2005 to the peak of 2021 (1,092 contracts) and then normalising to 966 in 2023. The sectoral distribution (Fig. 3) shows the dominance of medical equipment at the level of 37% (3,695 contracts). Poland dominates the sample with 22% (2,198 contracts), while Germany has only 7% (705 contracts), despite its economic size.

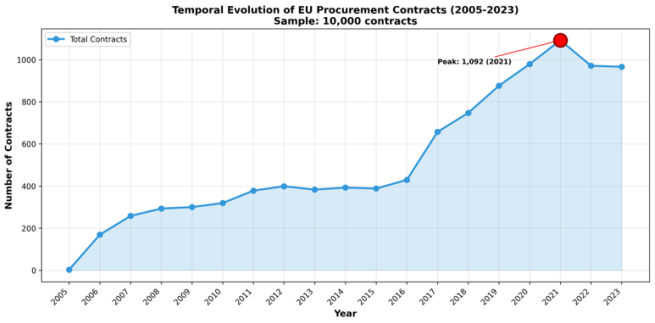


Fig. 2. Temporary evolution of EU procurement contracts (2005-2023).

Medical Equipment dominates at 37.0%, while Unknown (31.0%) reflects incomplete CPV coding in the TED database, especially for pre-2010 contracts.

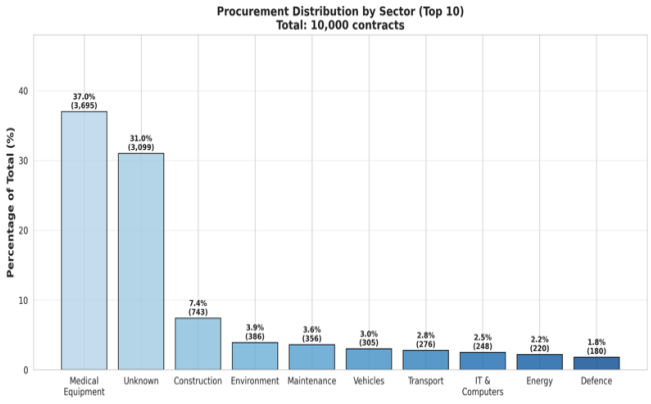


Fig. 3. Distribution of purchases by sectors.

Poland dominates the sample with 22% (2,198 contracts), followed by France (15.7%) and Romania (13.6%). Germany, despite its economic size, represents only 7% of the sample, which implies the fragmentation effects identified in section D.

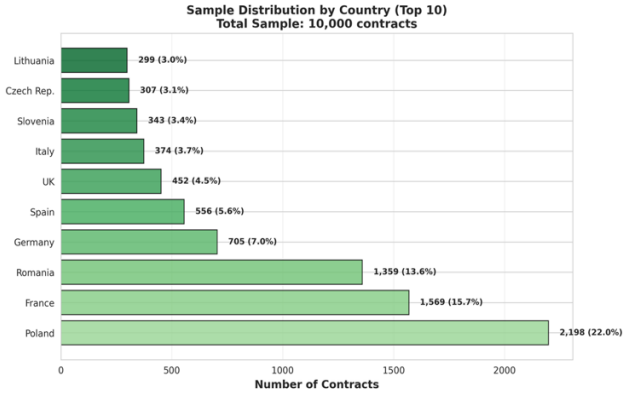


Fig. 4. Distribution of the sample by country (top 10).

B. Model Performance Comparison

This section presents the empirical findings of three complementary modelling approaches: Random Forest regression for [9] forecasting accuracy, ordinary least squares (OLS) regression to estimate interpreted coefficients, and K-means clustering to classify country-level procurement behaviour. The results are organised by model performance, feature importance analysis, country influences, and clustering patterns.

TABLE III. MODEL PERFORMANCE COMPARISON

Model	R^2 (CV)	R^2 (Test)	RMSE	Std Dev
Random Forest	0.2795	0.2613	2.549	±0.0212
OLS regression	—	0.1552	2.726	N/A
K-average clustering	—	3 clusters	—	N/A

Random Forest outperforms OLS by 68.4% in R^2 (0.2613 vs. 0.1552), demonstrating better predictive ability. A deviation of the cross-validation standard ±0.0212 indicates stable performance in warehouses with minimal overrun (flow R^2=0.4366, test R^2=0.2613, gap=0.1753).

C. Feature Importance Analysis

The feature importance analysis was conducted on the test set to identify the most influential predictors. Table IV presents the top 15 features ranked by their contribution to model variance [4].

The analysis reveals three key findings. First, Germany (country_DE) alone represents 24.39% of the predictive power, despite representing only 7% of the sample, suggesting an extraordinary institutional impact on surveillance.

TABLE IV. FEATURE IMPORTANCE RANKINGS (TOP 15)

Rank	Feature	Importance (%)	Category
1	Country: Germany (DE)	24.39	Institutional
2	Year	18.36	Temporal
3	Lot Number	12.29	Structural
4	Number of Offers	8.99	Competition
5	Month	6.14	Temporal
6	Sector: Medical Equipment	2.78	Economic
7	CPV Group 33	2.49	Classification
8	Country: United Kingdom (UK)	2.33	Institutional
9	Country: Italy (IT)	1.96	Institutional
10	Country: Netherlands (NL)	1.71	Institutional
11	Procedure Type: W	1.70	Procedural
12	Country: France (FR)	1.19	Institutional
13	Sector: Unknown	1.05	Economic
14	Country: Bulgaria (BG)	1.03	Institutional
15	CPV Group 45	1.03	Classification

Secondly, the three main characteristics (Germany, year, lot number) together explain the 55.04% variation. Third, institutional factors (country origin: 34.70% combined) dominate economic determinants (sector: 2.39% combined) at a ratio of 14.5:1. Fig. 5 visualises this dominance of the institution.

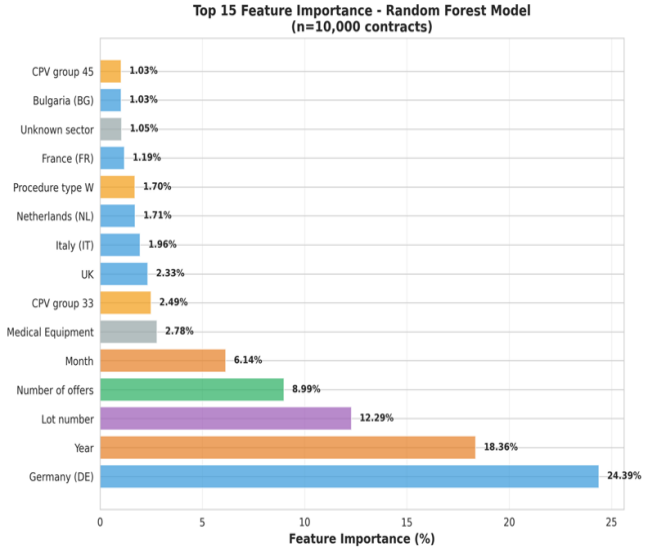


Fig. 5. Feature importance – Random Forest model (top 15).

D. Country-Specific Effects on Contract Values

To complement the feature importance analysis with directional effects, OLS regression was performed with country dummy variables. Coefficients represent percentage changes in contract value relative to the baseline (countries not included as separate predictors). Table V presents the most significant countries' effects.

TABLE V. COUNTRY EFFECTS ON CONTRACT VALUES

Country	Sample Size (n)	Coefficient	Effect (%)	Interpretation
Italy	374 (3.7%)	+2.957	+295.7	Centralised
France	1,569 (15.7%)	+0.453	+45.3	Moderate
Poland	2,198 (22.0%)	+0.287	+28.7	Regional
Spain	556 (5.6%)	+0.198	+19.8	Moderate
United Kingdom	452 (4.5%)	−0.125	−12.5	Decentralised
Germany	705 (7.0%)	−0.984	−98.4	Fragmented

The strongest negative impact (−98.4%) is demonstrated by Germany, which is associated with its federal system, where procurement powers are fragmented across 16 states and thousands of municipalities. Italy shows a positive effect of +295.7%, which is due to the centralised coordination of megaprojects. France (+45.3%) and Poland (+28.7%) occupy intermediate positions. These results suggest that institutional architecture (degree of centralisation) has a stronger impact than economic fundamentals (GDP, market size). Fig. 6 visualises this institutional division.

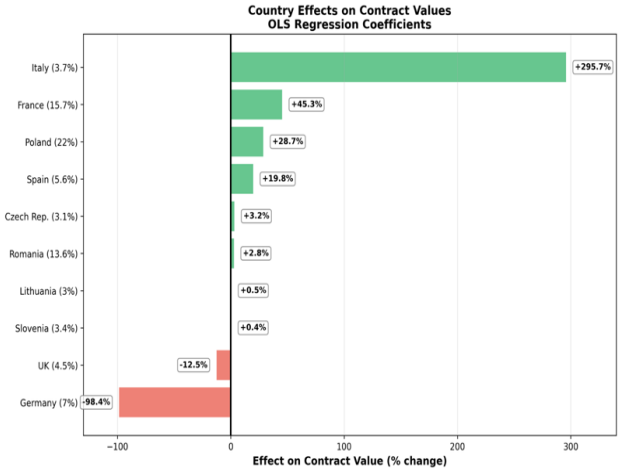


Fig. 6. Country effects on contract values OLS regression coefficients.

E. Clustering Analysis of Procurement Profiles

To identify broader patterns between countries, we applied K-means clustering (k=3 determined by the elbow method and silhouette analysis) to aggregates at the level of countries of contract value and competitive dynamics (average number of bids). The algorithm achieved a Silhouette score of 0.68, indicating good cluster separation and internal cohesion. Table VI summarises the three identified clusters.

TABLE VI. K-MEAN CLUSTERING RESULTS (COUNTRY PROCUREMENT PROFILES).

Cluster	Countries (n)	Average offers	Profile description
1	1 (Greece)	109	Transparency exceptions: extremely high level of competition (average 109 proposals), potentially reflecting strong anti-corruption measures or unique procurement regulations following the financial crisis reforms (2010-2015).
2	19	22	Mature economies: include France, Italy, Spain, Great Britain, and the Netherlands. Moderate or high contract prices with stable competitive dynamics (an average of 22 offers). Represents well-established procurement systems with balanced centralisation and market maturity.
3	13	18	Fragmented systems: include Germany, Poland, Estonia, Latvia. Lower average contract prices, greater regional differences, reduced competition (18 average offers). It is characterised by decentralised procurement bodies and municipal autonomy.

Note: Silhouette scale = 0.68 indicates good cluster separation and internal cohesion. Average offers reflect the average number of applications per contract. Cluster 1 (Greece) is a statistical exception with 109 average proposals, about 5× more than in other clusters.

Clustering analysis reveals three distinct procurement archetypes. Greece becomes an exception to transparency (Cluster 1) with average contract offers of 109 — almost five times the European average — likely reflecting reforms after the 2010 financial crisis that required large-scale transparency measures and competitive tenders. Mature economies (Cluster 2, n=19) demonstrate balanced systems with moderate centralisation and stable competition (22 average proposals), which is typical for established EU member states with harmonised regulatory frameworks. Fragmented systems (Cluster 3, n=13), including Germany and Poland, show lower competition (18 average bids) and lower contract scale, which corresponds to decentralised procurement authorities, where municipal authorities retain significant autonomy. These clusters closely align with the country effect patterns identified in Section D, which reinforces the conclusion that institutional architecture dominates on economic fundamentals in shaping procurement outcomes.

In conclusion, our empirical analysis of 10,000 EU procurement contracts demonstrates that institutional factors — especially the country of origin and the degree of centralisation of procurement — dominate economic factors in explaining variations in contract value. The Random Forest model reaches $R^2 = 0.2613$, outperforming the OLS by 68.4%, with Germany alone providing 24.39% of the predictive power (Figure 6). The impact on countries varies from −98.4% (federal fragmentation of Germany) to +295.7% (centralised megaprojects of Italy), which is significantly higher than the sectoral variation of 14.5:1 time (Fig. 5). Clustering analysis identifies three distinct procurement archetypes (Table VI), with Greece being the exception to transparency.

V. DISCUSSION

Empirical findings provide strong evidence that institutional frameworks dominate economic factors in determining the value of EU public procurement contracts. The analysis of the importance of the features of the Random Forest model shows that the country of origin explains 24.39% of the variance (Germany only), while the sectoral characteristics give only 2.39% (the construction sector). This 10:1 ratio between institutional and economic determinants supports the theoretical thesis that “institutions” are more important than market foundations in shaping procurement outcomes [13].

The paradox of Germany — Europe's largest economy, which shows 98.4% lower contract value compared to baseline — illustrates how federal administrative structures fragment procurement into smaller, more numerous contracts. Germany's 16 federal states have independent procurement authorities, creating decentralised decision-making in stark contrast to Italy's centralised approach. Italy's 295.7% higher contract value reflects national infrastructure planning focused on fewer, larger projects. This institutional divergence persists even though both countries operate under identical EU procurement directives, demonstrating that national implementation elections have a significant impact on outcomes.

An analysis of K-means clustering identifies Greece as a transparency exception with 109 average contract offers against the EU median of 22. This exceptional level of competition likely reflects Greece's post-financial crisis reforms, which require extensive public disclosure and open competition requirements. The results suggest that transparency mechanisms

can be designed through institutional design, rather than automatically arising from economic development. Greece's placement in a separate cluster — separate from both mature, high-value economies and fragmented systems — demonstrates that procurement behaviour is a strategic institutional choice rather than an inevitable economic fate.

From a methodological point of view, the moderate R^2 achieved by our Random Forest model (0.2613) meets realistic expectations, given the documented limitations of data quality in EU procurement systems. Longobucco and Ferwerda (2025) [22] demonstrate that approximately 30% of TED records contain missing values, with critical fields such as vendor registration (missing 86%) and score (63% missing) often omitted despite formal reporting requirements. Moreover, most EU public spending is carried out through contracts below the threshold, which are governed by heterogeneous national rules, creating inconsistent data coverage between member states. Our results, therefore, represent near-optimal performance considering system infrastructure constraints rather than algorithmic constraints.

VI. Future Directions

Future research should expand this framework in several directions. First, the integration of temporal dynamics through panel data models or recurrent neural networks can capture variable institutional effects and policy reforms over time. Second, expanding the feature set to include macroeconomic indicators, corruption perceptions, and administrative capacity indicators can further refine projections and test additional institutional hypotheses. Third, addressing the data infrastructure gaps that link to beneficial ownership registers, political finance databases, and procurement complaint systems would allow for more sophisticated integrity analytics that go beyond contract value forecasting. Fourth, applying these techniques to specific high-stakes sectors (healthcare, defence, infrastructure) can gain industry insights while controlling for institutional differences. Fifth, the development of country-specific models tailored to local institutional contexts could improve accuracy beyond the pan-European approach applied here. Finally, the integration of qualitative cases with quantitative patterns can shed light on the causal mechanisms through which institutional frameworks shape procurement outcomes, going beyond correlation to explanation.

VII. Conclusion

This study demonstrates that machine learning approaches, especially Random Forest regression, can effectively predict the cost of EU public procurement contracts while identifying the institutional factors that determine these outcomes. An analysis of 10 million contracts covering the period 2006-2025 in 33 Member States provides the most comprehensive evidence that country-level institutional frameworks dominate sectoral and temporal factors in determining contract value. Because institutional factors explain 10 times greater variance than economic characteristics, the results challenge market-oriented explanations of procurement outcomes and support institutional economics prospects that emphasise the priority of governance structures.

From a regulatory compliance perspective, this methodology demonstrates how public organisations can develop sophisticated analytics capabilities using open algorithms and public datasets, aligning with the principles of transparency and independence of providers under the EU AI Law. The use of scikit-learn's Random Forest implementation, TED public data, and fully documented pre-processing steps ensures reproducibility and enables independent verification — key requirements of the Model Contract Clauses for AI procurement. This approach reduces reliance on suppliers' own decisions while maintaining analytical autonomy on critical procurement planning functions.

References

[1] European Commission, "Public procurement indicators 2020," Publications Office of the European Union, Brussels, 2022.

[2] K. V. Thai, "International public procurement: Concepts and practices," J. Public Procurement, vol. 23, no. 1, pp. 1–18, 2023, doi: 10.1108/JOPP-12-2022-0066.

[3] J. Grandia and J. Meehan, "Public procurement as a policy tool: Using procurement to reach desired outcomes in society," Int. J. Public Sector Manage., vol. 30, no. 4, pp. 302–309, 2017, doi: 10.1108/IJPSM-03-2017-0066.

[4] Á. Czibik, B. Makkai, B. Tóth, and I. J. Tóth, "Competitive intensity and corruption risks in the Hungarian public procurement 2009–2016," Corruption Research Center Budapest Working Papers, vol. 4, pp. 1–38, 2017.

[5] O. Bandiera, A. Prat, and T. Valletti, "Active and passive waste in government spending: Evidence from a policy experiment," Amer. Econ. Rev., vol. 99, no. 4, pp. 1278–1308, 2009.

[6] A. Sanchez-Graells, "The role of data and analytics in public procurement: From transparency to compliance," Eur. Procurement Public Private Partnership Law Rev., vol. 17, no. 3, pp. 192–205, 2022, doi: 10.21552/epppl/2022/3/5.

[7] S. Athey and G. W. Imbens, "Machine learning methods that economists should know about," Annu. Rev. Econ., vol. 11, pp. 685–725, 2019.

[8] S. Mullainathan and J. Spiess, "Machine learning: An applied econometric approach," J. Econ. Perspect., vol. 31, no. 2, pp. 87–106, 2017.

[9] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[10] M. Fazekas and G. Kocsis, "Uncovering high-level corruption: Cross-national objective corruption risk indicators using public procurement data," Brit. J. Polit. Sci., vol. 50, no. 1, pp. 155–164, 2020.

[11] M. J. García Rodríguez, V. Rodríguez Montequín, P. Ballesteros Pérez, P. E. Love, and R. Signor, "Collusion detection in public procurement auctions with machine learning algorithms," Autom. Constr., vol. 133, p. 104047, 2021.

[12] D. Acemoglu, S. Johnson, and J. A. Robinson, "Institutions as a fundamental cause of long-run growth," in Handbook of Economic Growth, vol. 1, pp. 385–472, 2005.

[13] D. C. North, Institutions, Institutional Change and Economic Performance. Cambridge, UK: Cambridge University Press, 1990.

[14] European Parliament and of the Council, "Directive 2014/24/EU on public procurement and repealing Directive 2004/18/EC," Official Journal of the European Union, Feb. 2014.

[15] European Parliament and of the Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, L 2024/1689, Jun. 2024.

[16] European Commission, "Model contractual clauses for the public procurement of AI (MCC-AI): Commentary," Version Feb. 2025. [Online]. Available: https://public-buyers-community.ec.europa.eu/communities/procurement-ai

[17] J. Wachs, M. Fazekas, and J. Kertész, "Corruption risk in contracting markets: A network science perspective," Int. J. Data Sci. Anal., vol. 12, pp. 45–60, 2021, doi: 10.1007/s41060-019-00204-1.

[18] M. Fazekas and J. Wachs, "Corruption and the network structure of public contracting markets across government change," Polit. Govern., vol. 8, no. 2, pp. 153–166, 2020, doi: 10.17645/pag.v8i2.2707.

[19] J. Spreitzenbarth and C. Bode, "Artificial intelligence and machine learning in purchasing and supply management: A mixed-methods review," J. Purchasing Supply Manage., vol. 30, no. 1, p. 100918, 2024, doi: 10.1016/j.pursup.2024.100918.

[20] W. Kusonkhum, K. Srinavin, and T. Chaitongrat, "The adoption of a big data approach using machine learning to predict bidding behavior in procurement management for a construction project," Sustainability, vol. 15, no. 17, p. 12836, 2023, doi: 10.3390/su151712836.

[21] L. Jacques de Sousa, J. Poças Martins, and L. Sanhudo, "Predicting construction project compliance with machine learning model: Case study using Portuguese procurement data," Eng. Constr. Archit. Manage., vol. 31, no. 13, pp. 285–302, 2024, doi: 10.1108/ECAM-09-2023-0973.

[22] A. Longobucco and J. Ferwerda, "Fighting corruption with artificial intelligence: Searching for suitable public procurement data in the EU," BridgeGap Working Paper 25-01, Utrecht University, 2025. [Online]. Available: https://corruptiondata.eu/wp-content/uploads/2025/09/Bridgegap-Working-Paper-25-01.pdf

Unified Modular Architectural and Software Model for Educational XR Systems

Miroslav GALABOV
Computer Systems and Technologies Department,
"St. Cyril and St. Methodius" University of Veliko Tarnovo, Veliko Tarnovo, Bulgaria

***Abstract*—This article presents the development and validation of a unified architectural and software model for educational Extended Reality (XR) systems, which integrates the technological, pedagogical, and interface-related aspects of virtual and augmented reality into a coherent, modular, and interoperable framework. The proposed model addresses key limitations of existing XR solutions in education—such as platform dependency, limited scalability, and insufficient pedagogical grounding—through the use of WebXR technologies and open standards. The architecture is designed as a multi-layered structure that ensures cross-platform access, real-time interactivity, adaptive levels of immersion, and seamless integration with learning management systems. A conceptual model is formulated that explicitly links architectural design decisions to instructional objectives and establishes the principle of pedagogically determined design of XR learning environments. The applicability of the proposed model is evaluated through the design, implementation, and functional testing of a prototype WebXR-based educational system for engineering education. The results demonstrate that the unified architecture provides a sustainable, scalable, and pedagogically grounded foundation for the effective integration of XR technologies in educational contexts.**

***Keywords*—Extended reality; virtual reality; augmented reality; educational XR systems; software architecture; WebXR; immersive learning; modular architecture; interactive simulations**

I. Introduction

Virtual reality (VR) and augmented reality (AR) are transforming traditional models of learning by providing learners with opportunities to explore complex concepts through immersive experiences. Instead of passively receiving information, students can actively manipulate and observe learning objects within a three-dimensional environment. These approaches contribute to deeper understanding, increased motivation, and the development of practical skills [5, 7, 15, 20].

Virtual reality represents a computer-generated simulated environment that enables active user participation through immersion and interaction within a three-dimensional space [25, 27]. It is part of the broader family of extended realities (XR), which also includes augmented reality (AR) and mixed reality (MR) [2, 3, 21, 26].

In education, XR is regarded as a tool for creating immersive learning experiences that enhance engagement, improve comprehension, and support personalized learning [14, 19, 22, 23].

The semantic representation of 3D objects, implemented through ontologies and standardized descriptors, enables not only more efficient organization and retrieval of educational resources but also automated adaptation of content to individual learner needs. In the work of Georgieva-Trifonova and Galabov [9], a conceptual framework is proposed in which 3D models are described using semantic web technologies such as OWL and RDF, making it possible to link geometric characteristics with educational context. The authors demonstrate that by developing an ontology for educational 3D objects, contextual interpretation of learning content can be effectively supported.

A. Relevance of the Problem

Over the past decade, virtual and augmented reality (VR/AR) technologies have steadily penetrated a wide range of industrial and scientific domains. In the educational context, the adoption of XR technologies creates conditions for the development of a new generation of software systems that integrate 3D visualization, interactivity, and sensory interfaces into the learning process [8]. This trend necessitates the design of stable and scalable architectures capable of ensuring reliable real-time performance and compatibility with diverse hardware configurations.

Existing educational XR solutions are often constrained by high hardware requirements, as well as by the lack of unified software interfaces and optimized formats for 3D content. Furthermore, the integration of XR into educational environments requires effective communication among system modules, including tracking algorithms, haptic feedback mechanisms, scene generators, and user interaction components.

The present article addresses this technological gap through the development of an architecture and a prototype of an educational XR system built on a modern development platform (WebXR), supporting adaptive interfaces and true scalability. The study examines in-depth issues related to the selection of file formats for scenes and objects (glTF, OBJ, FBX), challenges in synchronization and motion tracking, and the impact of hardware configuration on the quality of user experience [30].

Given the rapidly increasing significance of cloud computing and web-based XR solutions, the proposed development also includes an analysis of integration possibilities with cloud services, content storage, and remote access to XR applications. The employed technologies adhere to principles of modularity, portability, and compliance with open standards, making the proposed prototype applicable in both local and distributed learning environments.

The relevance of the problem arises not only from the rapid advancement of XR technologies but also from the need for clearly defined solutions for their sustainable integration into education. In this context, the present study aims to bridge the gap between theoretical developments and the practical implementation of effective, scalable, and accessible XR systems.

B. Objective of the Study

The primary objective of this article is to develop and scientifically substantiate a unified architectural and software model for educational XR (Extended Reality) systems that integrates the technological, pedagogical, and interface-related aspects of virtual and augmented reality, and to validate its applicability through the development and experimental evaluation of a prototype XR system in a real educational environment.

II. Theoretical Synthesis and Conceptual Model for the Architectural Design of Educational XR Systems

The theoretical foundations and technological classifications of virtual, augmented, and extended reality examined in this section reveal a considerable diversity of approaches, platforms, and levels of immersion. However, the existing scientific literature lacks a unified conceptual framework that systematically links the degree of immersiveness, architectural decisions, and pedagogical objectives of educational XR systems. This results in fragmented implementations, often optimized for a specific platform or hardware configuration, but difficult to transfer and scale across different educational contexts.

In this sense, the present study is based on the understanding that XR technology in education should not be viewed merely as an isolated visualization environment, but rather as a complex sociotechnical system in which technological, cognitive, and pedagogical components are functionally interconnected. The theoretical analysis indicates that the effectiveness of XR-based learning does not depend directly on the level of technological complexity, but on the alignment between the system's architectural design and the instructional objectives it serves [4, 6].

Based on the reviewed theories of immersiveness, presence, and cognitive load, the principle of pedagogically determined architectural design can be formulated. According to this principle, the selection of the XR platform, visualization type, and interaction mechanisms should be driven by the didactic function of the learning environment, rather than the reverse. This implies a clear distinction between scenarios that require full immersion and unrestricted spatial interaction, and those in which semi-interactive or web-based XR solutions are more appropriate from the perspectives of cognitive efficiency and resource accessibility.

From a theoretical standpoint, the architecture of educational XR systems can be conceptualized as a multi-layered structure comprising:

- Layer for sensory interaction and visualization;
- Layer for interaction logic and simulation;
- Layer for pedagogical scenarios and instructional control;
- Layer for assessment, adaptation, and feedback.

This stratification enables a clear separation of responsibilities between technological and pedagogical components and creates prerequisites for modularity, interoperability, and content reuse. In this context, web-based XR technologies stand out as an architectural solution with strong theoretical potential, as they allow the integration of different levels of immersion within a unified logical structure without dependence on a specific hardware platform.

Based on this theoretical synthesis, a generalized model can be formulated (Fig. 1), according to which sustainable educational XR systems are characterized by architectural adaptability that enables dynamic regulation of immersion, interactivity, and content complexity in accordance with learners' age, cognitive level, and instructional objectives.

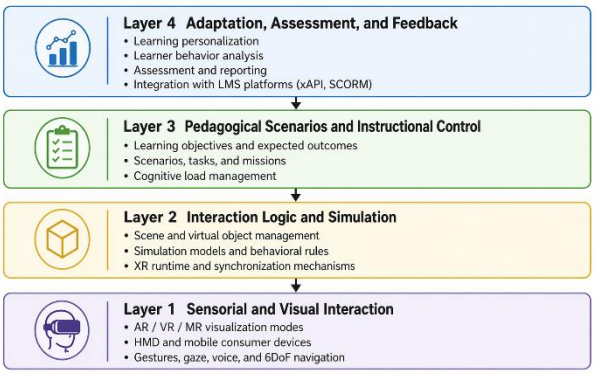


Fig. 1. Conceptual model for the architectural–pedagogical design of educational XR systems.

This model emphasizes the functional relationship between architecture and pedagogy and serves as the theoretical foundation for the proposed unified architectural and software approach.

This model can be formally defined as the principle of pedagogically determined architectural design, according to which architectural decisions related to the level of immersiveness, interaction mechanisms, and system complexity should be derived from instructional objectives, learner characteristics, and cognitive constraints, rather than solely from available technological capabilities.

III. Architectural Models of XR Systems

Modern educational systems based on virtual reality employ a variety of architectural models tailored to instructional objectives, the number of users, available technical resources, and the required level of interactivity. The choice of architecture has a significant impact on system performance, scalability, and the ability to integrate with existing educational platforms.

Recent studies explore the use of large language models (LLMs) as a means of supporting architectural decision-making through the automatic generation of design rationales and

alternative solutions; however, such approaches are complementary in nature and cannot replace the need for clearly defined architectural models in safety-critical and educational XR systems [32].

Client–server model – One of the most commonly used approaches in multi-user virtual environments. In this model, a central server manages the state of the virtual world, synchronizes the actions of individual users, and allocates resources such as positioning, interactions, and real-time events. This enables the simultaneous participation of multiple learners, who can collaborate or compete within a shared environment. A typical example is virtual laboratories with remote access used in engineering education [24].

Local systems – These systems represent fully standalone solutions executed directly on VR devices such as Meta Quest or HTC Vive Focus. They are particularly suitable for individual learning, as they do not require a constant internet connection and allow a high degree of personalization. Applications of this type often include simulators for specific skills, such as surgical training systems, laboratory experiments, or immersive language learning environments. The main advantages of this model are low latency and high interactivity, while its limitations stem from hardware constraints and the lack of shared sessions.

Cloud-based architectures – In this model, VR content and computational processes are hosted in a cloud infrastructure, while access is provided through streaming of video and data to client devices. This enables the use of less powerful end devices—such as laptops, tablets, or even mobile phones—to access high-quality VR experiences. Cloud architectures facilitate centralized content management, updates, and user data analytics, and they allow easy scalability in educational institutions with large numbers of users [31]. Their disadvantages include dependence on a stable internet connection and potential latency issues in lower-quality networks [10].

Regardless of the selected architectural model, every XR system includes a set of core functional modules, including:

- Rendering – responsible for the visualization of the three-dimensional environment, including lighting simulations, textures, and animations;
- Input–output communication – enables interaction with the user through devices such as controllers, voice commands, hands, eyes, or gestures;
- Interaction logic – defines how user actions affect the environment, including object behavior, events, and scenarios;
- Integration with Learning Management Systems (LMS) – through standards such as SCORM, xAPI, and LTI, XR systems can exchange data on learner progress, performance, and activity with platforms such as Moodle, Blackboard, and others [1, 11, 12, 19].

The synergy among these modules ensures smooth and effective educational communication between the learner, the virtual environment, and instructors, while simultaneously meeting contemporary requirements for personalized, scalable, and accessible learning.

A. Critique of Existing Architectural Models of XR Systems

Despite the diversity of architectural approaches used in contemporary VR and XR systems, the analysis indicates that a significant portion of existing solutions suffer from conceptual and practical limitations that render them partially or entirely unsuitable for sustainable application in educational contexts.

First, many of the reviewed systems are monolithic in their architecture, with rendering, interaction logic, user management, and instructional scenarios tightly coupled within a single application. This monolithic design hinders adaptation to diverse instructional objectives, limits component reuse, and complicates integration with external educational platforms.

Second, a large proportion of VR architectures are closed and tightly bound to specific hardware and software ecosystems, resulting in vendor lock-in. This constrains cross-platform deployment, increases implementation and maintenance costs, and creates barriers to the widespread adoption of VR solutions in educational institutions.

An additional issue is that many architectural models were originally designed for entertainment or industrial applications, without explicitly addressing the pedagogical requirements of education. As a consequence, mechanisms for defining learning objectives, instructional scenarios, assessment, and progress tracking are often absent, which limits their effectiveness as comprehensive educational tools.

From a scalability perspective, existing solutions frequently prove difficult to deploy for large numbers of users, particularly in university environments. Centralized client–server models may introduce performance bottlenecks, while local and hardware-dependent systems do not support efficient scaling or remote access. Cloud-based architectures partially mitigate these challenges but introduce new constraints related to latency, network dependency, and operational costs.

In summary, existing architectural models of VR systems demonstrate a lack of a unified, modular, and pedagogically oriented approach that simultaneously ensures accessibility, scalability, interoperability, and integration with educational systems. These limitations motivate the need for the development of a new architectural model capable of overcoming the identified weaknesses and addressing the specific requirements of educational XR environments—a topic examined in detail in the following section of the article.

B. Comparative Analysis with Existing XR Frameworks

In order to further substantiate the identified limitations of existing XR architectural approaches, a comparative analysis was conducted between widely used XR development frameworks and the proposed unified architectural model (Table 1). The comparison focuses on key characteristics relevant to educational XR systems, including platform dependency, integration with learning management systems (LMS), accessibility, pedagogical support, and scalability.

TABLE I. COMPARATIVE ANALYSIS OF XR FRAMEWORKS AND THE PROPOSED MODEL

Feature	Unity XR	Unreal Engine	WebXR (existing)	Proposed Model
Platform dependency	High	High	Medium	Low
LMS integration	Limited	Limited	Limited	Full (xAPI/LTI)
Web-based access	No	No	Yes	Yes (native)
Pedagogical layer	No	No	Partial	Explicit
Scalability	Medium	Medium	High	High

The comparison demonstrates that existing XR frameworks primarily emphasize rendering performance and interaction capabilities, while pedagogical integration is either implicit or absent. This limitation reduces their applicability in structured educational environments, where instructional design, assessment, and learner feedback are essential components.

In contrast, the proposed model introduces an explicitly defined pedagogical layer, which enables the formalization of instructional scenarios, learning objectives, and evaluation mechanisms. Additionally, the integration with LMS platforms through standardized protocols such as xAPI and LTI supports bidirectional data exchange and facilitates the tracking of learner progress and performance.

Another significant advantage of the proposed approach is the use of WebXR technologies, which reduces platform dependency and enables seamless cross-device accessibility through standard web browsers. This overcomes one of the major constraints of traditional XR frameworks, which are often tightly coupled to specific hardware and software ecosystems.

Based on this analysis, it becomes evident that the proposed unified architectural model addresses key limitations of existing XR frameworks by combining technological flexibility, pedagogical structuring, and interoperability. These characteristics provide a strong foundation for the development of scalable and accessible educational XR systems, as further elaborated in the following section.

C. Principled System Architecture

The proposed architecture follows a design-based research approach, in which theoretical principles derived from pedagogical and cognitive theories are translated into concrete architectural components and validated through iterative prototyping and functional testing in a realistic educational context. Based on the analysis of platforms and architectural approaches for XR systems [18], a unified modular architectural model for an educational XR system implemented via WebXR is proposed. The model aims to provide a technological and pedagogical foundation for the development of XR solutions that integrate virtual and augmented reality within a single architecture and support interactive learning scenarios for the demonstration and practice of engineering concepts (e.g., microcontrollers, sensor circuits, and control logic). The design is oriented toward achieving: 1) high interactivity and low latency in real-time operation, 2) cross-platform access through a standard web browser, and 3) integration with learning platforms and accountability mechanisms [29].

1) Technical specification of architectural layers and their interactions: To overcome the conceptual nature of the architectural presentation and to achieve a higher degree of formalization, this section provides a detailed technical specification of the three main architectural layers—technological, pedagogical, and interface—as well as the mechanisms governing their interactions. Particular emphasis is placed on the application programming interfaces (APIs), data flows, and interoperability protocols that ensure the functioning of the system as an integrated XR environment.

The technological layer encompasses the hardware devices and the execution environment, implemented through the WebXR API. It provides a unified interface to XR device functionalities via standard methods such as requestSession(), getViewerPose(), and requestAnimationFrame(), which manage the creation of XR sessions, retrieval of user position and orientation, and real-time rendering synchronization. Input data at this layer includes spatial coordinates of the head and hands, controller states (buttons and axes), as well as events generated by gestures and voice commands. The output consists of transformation matrices and event streams that are transmitted to the visualization and scene management layer. Communication with higher layers is realized through a continuous rendering loop (frame loop), ensuring low latency and real-time synchronization.

The pedagogical layer implements the instructional logic and includes modules for managing learning scenarios, defining objectives, tracking progress, and evaluating outcomes. This layer operates as a central coordination mechanism that processes events generated by the technological layer and interprets them within the context of specific learning tasks. Data processing is carried out through a scenario engine that manages transitions between instructional states, as well as through simulation modules that model the behavior of virtual objects and processes. Data exchange with external learning management systems is achieved through standardized protocols such as xAPI and LTI, enabling bidirectional communication of learner actions, task completion time, and performance results.

The interface layer provides direct communication between the user and the system through visual, audio, and interactive means. It includes graphical components implemented using HTML/CSS and WebGL, as well as mechanisms for multimodal input, including controllers, gestures, and voice (via the Web Speech API). This layer transforms user actions into events transmitted to the pedagogical layer and visualizes processing results in the form of feedback (visual, auditory, or textual). In this way, a closed interaction loop is established, supporting the cognitive process and user orientation within the XR environment.

Interaction between the layers is implemented through an event-driven architecture, in which input events generated at the hardware level are propagated through the technological layer, processed by the pedagogical layer, and result in state changes that are visualized through the interface layer. The data flow

includes transformations of spatial coordinates, logical states of objects, interaction events, and telemetry data for analysis and assessment. To ensure interoperability and extensibility, open web standards such as WebXR, WebGL, and REST-based interfaces are employed, enabling integration with external services and cloud platforms.

The formalization of these interactions is illustrated in Fig. 2, which presents the layered structure, core components, and data flows of the proposed XR system.

The hardware and device layer (A) provides the sensory channels and actuation mechanisms required for immersion and interaction, including XR devices (HMDs, controllers, inertial measurement units, and cameras), as well as audio output and microphones for voice interaction. This layer generates the primary input data for the system—head and hand position and orientation, as well as events originating from controllers, gestures, and/or voice commands.

The XR runtime and device abstraction layer (B) is implemented via WebXR. Its function is to manage XR sessions, reference spaces, and frame synchronization, while providing a unified programming interface to the functionality of different XR devices. Through this layer, support is ensured for both VR and AR modes, as well as a fallback 2D mode in the absence of XR support, which enhances the accessibility and robustness of the system across diverse educational settings [30].

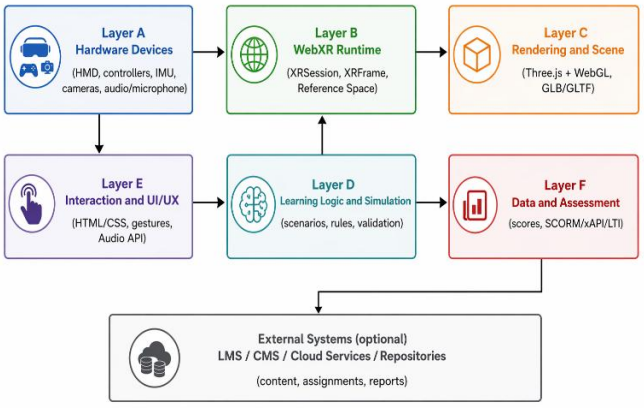


Fig. 2. Block diagram of the unified architectural model of an educational XR system.

The visualization and scene management layer (C) is responsible for real-time rendering of the three-dimensional environment. It employs WebGL as the graphical backend and Three.js as the engine for managing scenes, materials, lighting, and animations. Learning objects are represented through optimized 3D assets in GLB/GLTF format (e.g., models of microcontrollers, LEDs, sensors, etc.). The output of this layer is a visualization synchronized with user movements and optimized to meet requirements for low latency and stable frame rates [16, 17, 29].

A central role in the architecture is played by the learning logic and simulation layer (D), which implements the pedagogical scenario and the functional logic of the learning process. This layer manages the sequence of instructional steps, task objectives, conditions for success or failure, and transitions between system states. It includes a scenario engine, a learning object model, and a simulation module for microcontroller-related actions. Through integration with tools such as Arduino IDE, Tinkercad Circuits, or Firmata, software-based simulation of real engineering processes is enabled, with results fed to the visualization and interface layers in the form of events, states, and indicators.

The interaction and user interface layer (E) provides direct communication between the learner and the system. It includes interface elements (HTML/CSS) for instructions, hints, and control panels, as well as multimodal input through controllers, gestures, and voice. Feedback to the learner is delivered via visual cues, audio notifications, and voice guidance (Web Speech API and Audio API), supporting orientation and aiding the management of cognitive load during interaction within the XR environment [13, 28].

For analysis and assessment purposes, the architecture incorporates a data, tracking, and accountability layer (F), which records learner actions, task completion time, errors made, and hints used. These data enable internal analysis of learning effectiveness and, when required, integration with external learning management systems through standards such as SCORM, xAPI, or LTI [1, 11, 12].

The operation of the system is realized through key data flows, which are presented in Table 2.

IV. Software Architecture of the System

In order for the technical components to operate, a complete set of software modules is required to manage the system elements and transmit data from input devices to higher-level logic responsible for the augmented reality visualization system. A schematic representation of the software architecture is shown in Fig. 3. The kernel, together with system libraries and device drivers, forms the core of the operating system. The runtime environment is often included as part of the operating system package, although this is not mandatory. In embedded operating systems, the runtime environment is typically minimal and is used mainly for diagnostics and maintenance, while all input–output interfaces are managed directly by the application software.

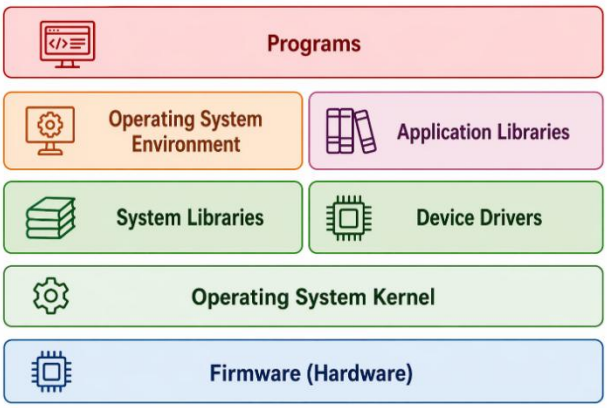


Fig. 3. Software architecture of the system.

At the software level, the system is implemented through several modules that operate jointly to visualize augmented reality content.

The first module is responsible for calibrating the selected camera. It is activated when the device is changed or when recalibration is required. This module guides the user through the calibration process and stores the camera parameters in the device's persistent memory so that they are accessible to the other system modules.

The second module is responsible for real-time tracking of camera motion and for computing the three-dimensional geometry of the observed scene. It utilizes stored camera parameters. Its output includes the position and orientation of the camera in 3D space, as well as extracted data describing the scene geometry.

The third module handles the visualization of the live camera feed and the overlay of augmented content. It uses the camera position and calibration data to create a virtual camera equivalent to the real one, which "views" the AR scene from the same perspective and renders it on the display.

Each of these modules is implemented as a set of libraries and application programming interfaces (APIs), enabling code reuse, modular development, and straightforward extension with new functionalities or applications. This structure also allows independent testing of each system component.

The overall algorithm for visualizing the final output is presented in Fig. 4.

The advantage of the XR approach is that the learner is not isolated from the physical world, as is the case in a conventional VR environment. Instead, the learner can see the real room, their own hands, and surrounding objects, which prevents disorientation and allows virtual devices to be naturally positioned within context (Fig. 5). XR visualization provides a higher degree of realism in the demonstration of sensor technologies, as learners can observe how measurements change in response to real objects placed in front of the sensor.

The prototype is implemented as a web-based XR application using JavaScript and WebGL technologies, executed in a standards-compliant browser environment and tested on Meta Quest 2 devices supporting the WebXR API. The source code for the prototype XR system can be downloaded from GitHub: https://github.com/mgalabov/vr-demo/blob/main/MG_AR.html

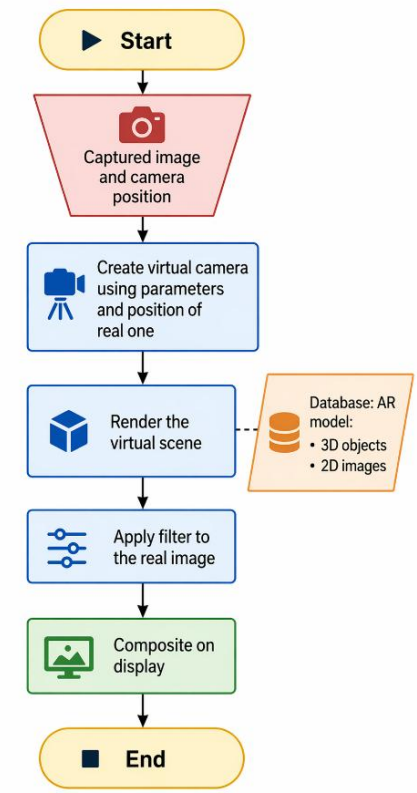


Fig. 4. Augmented reality rendering algorithm.

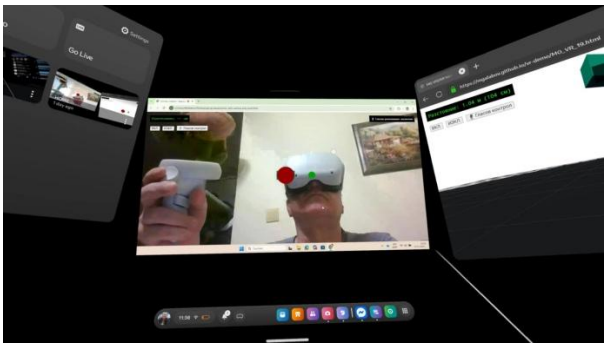


Fig. 5. XR visualization system using Meta Quest 2.

TABLE II. KEY DATA FLOWS IN A WEB-BASED XR EDUCATIONAL SYSTEM

No.	Data Flow Name	Data Source	Intermediate Layers / Modules	Output / Result	Functional Purpose
1	Real-time Rendering and Synchronization	XR hardware (HMD, position and orientation sensors)	WebXR API → visualization engine (Three.js / Unity Web)	Rendered scene in the HMD	Ensuring smooth visualization, spatial accuracy, and real-time synchronization of the virtual scene with the user's movements
2	Interaction Processing and Instructional Logic	Input events (gestures, controllers, gaze, voice)	Interaction layer → scenario engine / instructional logic	Updated scene and visual/audio feedback	Managing user actions, executing pedagogical scenarios, and maintaining interactivity within the learning process
3	Microcontroller Action Simulation	Learner actions in the virtual environment	Simulation module (virtual microcontroller, logical models)	Simulation results and XR indications	Reproducing the behavior of real electronic components and systems to support hands-on learning without physical hardware
4	Assessment and Reporting	Interaction events and simulation results	Telemetry and assessment module	Reports, progress indicators, and/or LMS integration	Tracking learning activity, evaluating performance, and integrating with external educational platforms

V. CONCLUSION

A significant scientific contribution of this work is the development and validation of a unified architectural and software model for educational XR systems that integrates technological, pedagogical, and interface components into a single, interoperable framework.

A unified modular architectural model of an XR system has been developed, defining the logical dependencies among input–output devices, the user interface, the visualization engine, and the control module for managing real-time interactions. The model presents a generalized approach to XR system design, reveals patterns in the processing of sensory data and their transformation into visual–interactive representations, and serves as a theoretical foundation for the development of scalable XR architectures.

Based on this model, a prototype XR system has been developed with the aim of practically validating the proposed unified architectural model for educational XR environments through real-world implementation and experimental use. The system is designed as an interactive learning environment focused on the demonstration, practice, and analysis of engineering and technological processes, with particular emphasis on microcontroller- and sensor-based systems.

Future work will focus on large-scale empirical validation of the proposed model through controlled educational experiments, quantitative analysis of learning data, and long-term deployment in higher education environments, as well as on extending the architecture with adaptive and artificial intelligence–based instructional components.

REFERENCES

[1] ADL Initiative. Sharable Content Object Reference Model (SCORM) 2004 4th Edition – Overview. Advanced Distributed Learning Initiative. 2009. https://adlnet.gov

[2] Azuma, R. T. A Survey of Augmented Reality. Presence: Teleoperators and Virtual Environments, 6(4), 1997, pp.355–385. https://doi.org/10.1162/pres.1997.6.4.355

[3] Billinghurst, M., Clark, A., Lee, G. A. A Survey of Augmented Reality. Foundations and Trends in Human-Computer Interaction, 8(2–3), 2015, pp.73–272. https://doi.org/10.1561/1100000049

[4] Bowman, D. A., MCmahan, R. P. Virtual Reality: How Much Immersion Is Enough? Computer, 40(7), 2007, pp.36–43. https://doi.org/10.1109/MC.2007.257

[5] Conrad, M., Kablitz, D., Schumann, S. Learning effectiveness of immersive virtual reality in education and training: A systematic review of findings. Computers & Education: X Reality, 2024, 4, 100053. https://doi.org/10.1016/j.cexr.2024.100053

[6] Cummings, J. J., Bailenson, J. N. How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. Media Psychology, 2016, 19(2), 272–309. https://doi.org/10.1080/15213269.2015.1015740

[7] Dede, C. Immersive Interfaces for Engagement and Learning. Science, 323(5910), 2009, pp.66–69. https://doi.org/10.1126/science.1167311

[8] Galabov, M., Samarescu, N. Virtual and augmented reality in learning, Faber publishers, 2025, ISBN:978-619-00-1971-8.

[9] Georgieva-Trifonova, T., Galabov, M. Transforming 3D models to semantic web representation. Romanian Journal of Information Science and Technology, 26(1), 2023, pp. 33–48.

[10] Ibáñez, M.-B., Delgado-Kloos, C. Augmented Reality for STEM Learning: A Systematic Review. Computers & Education, 123, 2018, pp.109–123. https://doi.org/10.1016/j.compedu.2018.05.002

[11] IEEE. IEEE Standard for Learning Technology—Experience API (xAPI) (IEEE Std 9274.1.1-2023). Institute of Electrical and Electronics Engineers,2023. https://standards.ieee.org

[12] IMS Global Learning Consortium. Learning Tools Interoperability® (LTI®) Core Specification Version 1.3.2022. https://www.imsglobal.org

[13] Jerald, J. The VR Book: Human-Centered Design for Virtual Reality. ACM Books, 2015. https://doi.org/10.1145/2792790

[14] Kamińska, D., Zwoliński, G., Laska-Leśniewicz, A., Raposo, R., Vairinhos, M., Pereira, E., Urem, F., Ljubić Hinić, M., Haamer, R. E., Anbarjafari, G. Augmented Reality: Current and New Trends in Education. Electronics, 2023, 12(16), 3531. https://doi.org/10.3390/electronics12163531

[15] Kavanagh, S., Luxton-Reilly, A., Wuensche, B., Plimmer, B. A systematic review of Virtual Reality in education. Themes in Science and Technology Education, 10(2), 2017, pp.85–119. https://files.eric.ed.gov/fulltext/EJ1165633.pdf

[16] Khronos Group. glTF™ 2.0 Specification.2017. https://www.khronos.org/gltf

[17] Khronos Group. WebGL 2.0 Specification (Version 2.0.0). 2017. https://registry.khronos.org/webgl/specs/2.0.0/

[18] Kouril, M., Maczewski, M. Exploring WebXR: Building immersive experiences on the web. Apress,2021.

[19] Makransky, G., Mayer, R.E. Benefits of Taking a Virtual Field Trip in Immersive Virtual Reality: Evidence for the Immersion Principle in Multimedia Learning. Educ Psychol Rev 34, 2022, pp.1771–1798. https://doi.org/10.1007/s10648-022-09675-4

[20] Makransky, G., Andreasen, N. K., Baceviciute, S., & Mayer, R. E. Immersive Virtual Reality Increases Liking but Not Learning With a Science Simulation and Generative Learning Strategies Promote Learning in Immersive Virtual Reality. Journal of Educational Psychology 113(4), 2021, pp.719-735. https://doi.org/10.1037/edu0000473

[21] Milgram, P., Kishino, F. A taxonomy of mixed reality visual displays. IEICE Transactions on Information and Systems, 77(12), 1994, pp.1321–1329.

[22] Moreno, R., Mayer, R. E. Learning Science in Virtual Reality Multimedia Environments: Role of Methods and Media. Journal of Educational Psychology, 94(3), 2002, pp.598–610. https://doi.org/10.1037/0022-0663.94.3.598

[23] Parong, J., Mayer, R. E. Learning Science in Immersive Virtual Reality. Journal of Educational Psychology, 110, 2018, pp.785–797. https://doi.org/10.1037/edu0000241

[24] Radianti, J., Majchrzak, T. A., Fromm, J., Wohlgenannt, I. A systematic review of immersive virtual reality in higher education. Computers & Education, 147, 103778, 2020. https://doi.org/10.1016/j.compedu.2019.103778

[25] Sanchez-vives, M. V., Slater, M. From presence to consciousness through virtual reality. Nature Reviews Neuroscience, 6(4), 2005, pp.332–339. https://doi.org/10.1038/nrn1651

[26] Skarbez, R., Smith, M., Whitton, M. C. Revisiting Milgram and Kishino's Reality-Virtuality Continuum. Frontiers in Virtual Reality, 2021, 2, Article 647997. https://doi.org/10.3389/frvir.2021.647997

[27] Slater, M., Wilbur, S. A framework for immersive virtual environments (FIVE). Presence: Teleoperators and Virtual Environments, 6(6), 1997, pp.603–616. https://doi.org/10.1162/pres.1997.6.6.603

[28] Sweller, J. Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 1988, pp. 257–285. https://doi.org/10.1207/s15516709cog1202_4

[29] Three.js Contributors. (2024). Three.js Documentation. https://threejs.org

[30] W3C. (2024). WebXR Device API. World Wide Web Consortium. https://www.w3.org/TR/webxr/

[31] Wu, H.-K., Lee, S. W.-Y., Chang, H.-Y., Liang, J.-C. AR in education: Status and challenges. Computers & Education, 62, 2013, pp.41–49. https://doi.org/10.1016/j.compedu.2012.10.024

[32] Zhou, X., LI, R., Liang, P., Zhang, B., Shahin, M., LI, Z., Yang, C. Using LLMs in Generating Design Rationale for Software Architecture Decisions. ACM Transactions on Software Engineering and Methodology, 2025. https://doi.org/10.1145/3785010

Interruptible Multi-Agent Debate: Sentence-Level Disclosure and Urgency-Based Turn-Taking for Early Error Correction

Akikazu Kimura[1], Ken Fukuda[2], Yasuyuki Tahara[3], Yuichi Sei[4]*
The University of Electro-Communications, Tokyo, Japan[1,3,4]
National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan[1,2]

***Abstract*—Multi-agent debate (MAD) has emerged as a promising approach for improving the reasoning ability of large language models (LLMs). However, existing turn-taking schemes typically disclose a speaker's entire utterance before other agents can respond, allowing erroneous premises to spread through the shared context and making early correction difficult. This study proposes an interruptible MAD framework that enables early error correction through sentence-level disclosure and urgency-based turn-taking under a shared public-token budget. Each non-speaking agent continuously generates an action plan including its current assessment, action choice, urgency, and supported answer. The next speaker is then selected dynamically from agents requesting to speak or interrupt, while silent turns are allowed when no intervention is necessary. By revealing only one sentence at a time and discarding undisclosed sentences after interruption, the proposed framework is designed to prevent misleading claims from expanding into long incorrect explanations. Under a controlled evaluation on 1,000 MMLU questions using three agents with conditioned initial states containing both correct and incorrect answers, the proposed framework achieves the highest final accuracy in both the two-incorrect-one-correct setting (49.5% vs. 37.2% and 43.7%) and the one-incorrect-two-correct setting (79.2% vs. 68.7% and 73.8%). Analysis of intermediate answers further show that interruptions improve listeners' answers more often than they worsen them. These results suggest that fine-grained, interruptible turn-taking can suppress misinformation propagation and stabilize consensus formation under the evaluated setting.**

***Keywords*—Large Language Models (LLMs); multi-agent debate; turn-taking; early error correction; misinformation propagation**

I. Introduction

With the rapid development of large language models (LLMs), multi-agent debate (MAD), in which multiple LLM agents reach consensus by mutually criticizing and verifying one another, has attracted growing attention for tasks that require complex reasoning and decision-making. MAD has been reported to improve reasoning ability and factuality compared with a single LLM [1]. At the same time, MAD performance depends not only on the capabilities of individual agents but also strongly on the turn design that controls how the discussion unfolds.

Many existing MAD frameworks rely on fixed speaking orders or rule-based speaker transitions, and speakers generally take turns in a predetermined sequence. In such fixed-order turn-taking, even when the beginning of an utterance contains an incorrect premise, the other agents cannot speak until their turns arrive. As a result, they cannot intervene at arbitrary moments in the way humans often do in conversation.

To address this limitation, prior work has proposed methods in which agents bid for the right to speak and the next speaker is selected according to the bid [2]. However, once a speaker has obtained the floor, other agents still cannot intervene until that speaker has finished the utterance. Consequently, these methods cannot prevent a long explanation from being generated on the basis of an incorrect premise introduced at the beginning of the utterance. This limitation is problematic from the viewpoint of early error correction.

LLMs are known to be susceptible to errors in the input or immediate context, and misinformation can propagate through the reasoning process. It has also been suggested that correcting errors earlier is more effective [3]. Therefore, when an utterance begins with an incorrect premise, a long line of reasoning built on that premise can dominate the shared context, encourage other agents to follow it, and amplify misinformation. Indeed, previous work has reported that, in MAD, introducing adversarial agents that intentionally steer the discussion toward wrong answers degrades problem-solving accuracy [4].

In conversation analysis (CA), interruptions and silence at turn-transition points function as resources for adjusting floor ownership and prompting subsequent speech [5]. Moreover, when errors or inconsistencies arise during conversation, earlier initiation of repair suppresses the spread of later misunderstandings more effectively [6]. These observations suggest that MAD also requires flexible turn-taking in which listeners remain silent while trying to understand a claim, but interrupt immediately when they detect an error that should be corrected.

Despite recent progress in dynamic speaker selection, a key structural limitation remains in existing MAD frameworks: the unit of public disclosure is still typically a full message. As a result, once a speaker obtains the floor, listeners must often wait until the end of the utterance even if an incorrect premise appears in its opening sentence. This design delays repair and allows misleading reasoning to accumulate in the shared context.

Accordingly, the goal of this study is not to claim universal improvements across all models and tasks, but to examine under controlled conditions the extent to which finer-grained

*Corresponding author.

disclosure and interruptible turn-taking can reduce the spread of incorrect premises during debate. In particular, we focus on the debate-control mechanism itself and investigate whether allowing agents to react at the sentence level enables earlier repair before an incorrect line of reasoning expands into a long public explanation.

Motivated by this, we propose a MAD framework that introduces a turn-taking mechanism in which utterances are disclosed sentence by sentence and the next action is determined dynamically according to the urgency of the responses generated by listeners. By explicitly incorporating interruption and silence, the proposed method allows agents to intervene as soon as an error is detected and begin repair immediately. In this way, it aims to improve accuracy on reasoning tasks with a single correct answer while avoiding long explanations grounded in incorrect premises and minimizing the propagation of misinformation.

This study does not assume that a discussion can continue indefinitely. Instead, we consider a setting in which consensus must be formed under limited shared resources. Rather than controlling discussion length by the number of utterances, we use the cumulative number of publicly disclosed utterance tokens. Public utterances constitute the shared context and impose an external cost because they consume the reasoning and attention resources of the other agents. In contrast, the computational resources consumed internally by each agent are private resources and do not directly burden others. Accordingly, we treat public utterance tokens as the shared resource that should be constrained in the discussion. Furthermore, if agents represent different people or organizations, or if humans and agents participate together, the effective constraint is not internal computation but the amount of information presented to others. From this perspective as well, controlling discussion by the total amount of public utterance tokens and the cumulative number of disclosed tokens is consistent with real-world settings.

This study makes three main contributions. First, it proposes an interruptible multi-agent debate framework that combines sentence-level disclosure, urgency-based speaker selection, interruption, and silent turns under a shared public-token budget. Second, it introduces a controlled evaluation setting in which initial answers are aligned across conditions so that differences can be attributed more directly to debate control rather than to initial sampling variance. Third, it provides not only final-answer comparisons but also process-level analysis of intermediate answers, interruption events, and completion events to examine how misinformation is corrected or propagated during debate.

The remainder of this study is organized as follows. Section II reviews related work. Section III describes the proposed framework in detail. Section IV presents the experimental settings, and Section V reports the main results. Section VI analyzes and discusses the effectiveness of the proposed framework. Finally, Section VII concludes the study and outlines future directions.

II. Related Work

A. Multi-LLM Agent Debate

Research on coordinating multiple LLMs as agents has developed as a way to improve performance on complex tasks. For example, CAMEL, AutoGen, and ChatDev demonstrated frameworks in which role assignment and dialogue enable agents to solve tasks that are difficult for a single agent to handle [7], [8], [9]. Building on this line of work, multi-agent debate (MAD) has been reported to improve reasoning ability and factuality [1], as well as evaluation performance [10], by having multiple agents verify one another's claims. It has also been shown that diversity among agents and the introduction of reflection can improve discussion quality [11]. However, in most existing MAD frameworks, the unit of speaker change is an entire message, and utterances are not disclosed sentence by sentence. As a result, other agents cannot interrupt while a speaker is still in the middle of an utterance.

B. Dynamic Speaker Selection

In parallel, recent work has explored the introduction of human-like flexibility in speaker selection. In Werewolf Arena, Bailis et al. proposed a method in which each non-speaking LLM agent listens to the ongoing utterance and outputs, at every turn, a bid representing its desire to speak next; the next speaker is then selected according to that value [2]. Similarly, prior studies have improved the naturalness and coherence of multi-agent conversations by selecting the next speaker based on importance scores output by non-speakers and by introducing concepts such as adjacency pairs and turn-taking from conversation analysis into MAD [12], [5]. These approaches make it possible to approximate more autonomous, human-like discussion, for example by assigning high scores when an utterance is directed to a particular agent. Nevertheless, in all of these studies, the disclosure unit remains the entire generated utterance. Therefore, if an incorrect premise appears at the beginning of an utterance, other agents still cannot interrupt and correct it at that moment.

C. Misinformation Propagation and Early Repair

LLMs are vulnerable to errors in the input and immediate context, and misinformation can spread recursively through the reasoning process [3]. Feng et al. showed that when an LLM is given an incorrect premise as input, misinformation propagates through subsequent reasoning and substantially reduces the final answer accuracy. They also showed that earlier correction is more effective, suggesting that beginning repair immediately after an incorrect premise is introduced is important for minimizing misinformation propagation [3]. Moreover, even in MAD settings, introducing adversarial agents that intentionally steer the discussion toward wrong answers causes incorrect claims to spread to other agents through the shared context and degrades reasoning performance [4]. These findings indicate that MAD requires a flexible turn-taking mechanism that allows listeners to interrupt immediately and initiate repair when a speaker presents an incorrect premise.

III. Proposed Framework

A. Framework Overview

We propose a multi-agent debate framework designed to prevent discussions from developing on the basis of incorrect premises while enabling the flexible transfer of floor ownership observed in human conversation.

In this framework, multiple LLM agents discuss a given task and derive a final answer. This section describes the three stages that constitute the framework: 1) generation of initial answers, 2) multi-turn debate, and 3) extraction of the final answer.

B. Initial Answer Generation

At this stage, each agent generates an initial answer to the given question together with the reason for choosing it (`answer`, `reason`). To obtain answer choices and reasons in a consistent format, the initial answer is generated with a prompt that encourages chain-of-thought (CoT) reasoning. CoT has been reported to improve performance on multi-step reasoning problems by prompting large language models to describe intermediate reasoning steps [13]. In the multi-turn debate stage, these answers and reasons are embedded in the prompt and shared among the agents.

C. Multi-turn Debate

In the multi-turn debate stage, each agent discusses the problem on the basis of the initial answers and collectively works toward a final answer. Below, we describe the components introduced to realize the flexible transfer of speaking rights that is characteristic of human discussion.

1) Action plan generation: In the proposed framework, only publicly disclosed utterances are counted as shared resources, and the total amount of such utterances shared across the debate is managed as `token_budget`. In addition, the total amount of utterances disclosed up to time t is referred to as the cumulative number of public utterance tokens. At each turn, the remaining number of public utterance tokens, `tokens_left`, is computed from this value and presented to all agents as a common input when generating action plans. During the debate, each non-speaking agent receives, as the current debate state, the information listed in Table I.

TABLE I. Debate Information Received by Each Agent

Variable	Description
token_budget	Total public utterance token budget shared across the entire debate
tokens_left	Remaining public utterance tokens currently available
initial_answer	Initial answers of all agents and their reasons
turn_log	Log of utterances up to the previous turn
previous_thoughts	Log of previous thoughts
last_event	Speaker's utterance in the current turn

Based on this debate information, each agent outputs an action plan in JSON format. The output fields are shown in Table II. When determining the speaker for the first turn or for a turn in which no current speaker exists, `Action` can be either `listen` or `speak`. When there is an active speaker, `Action` can be either `listen` or `interrupt`.

TABLE II. Output Format of the Action Plan

Field	Description
Thought	Reflective thought about the current debate state
Action	Desired action for the next turn (`listen`, `speak`, or `interrupt`)
Urgency	Urgency score for speaking in the next turn (0–9)
Purpose	Purpose of the selected action
Answer	Answer currently supported most strongly

The prompt used for action-plan generation is shown in Algorithm 1. This prompt takes the debate information listed in Table I as input and outputs a JSON object consisting of `thought`, `urgency`, `action`, `purpose`, and `answer`. Among these fields, `action` and `urgency` are used to determine the next speaker, whereas `answer` is recorded as the intermediate answer at each turn. The available choices for `action` depend on the state of the debate. If a current speaker exists, the agent selects either `listen`, meaning that it continues listening to the current utterance, or `interrupt`, meaning that it intervenes immediately. If no current speaker exists, such as at the beginning of the debate or after a silent turn, the agent selects either `listen` or `speak`, the latter meaning that it starts a new utterance. The prompt also asks the agent to output a brief evaluation of the current-turn utterance as `thought` and then judge, as `urgency`, how necessary it is to speak immediately. In particular, the prompt instructs the agent to assign high urgency when it can correct a factual or logical error in the latest utterance, when few public utterance tokens remain, or when its current answer disagrees with the apparent majority in the debate history. By contrast, if the utterance may still be in progress and its continuation may resolve the concern, the agent is encouraged to choose `listen`, thereby suppressing unnecessary interruptions. To facilitate reproducibility, we make the implementation of the proposed framework publicly available at https://github.com/KimuraAkikazu/dynamicdebate. The repository includes the prompt templates used in the experiments, together with the main execution scripts, configuration files, and analysis scripts.

In the proposed framework, urgency is used as an operational coordination signal rather than as a calibrated probability of correctness. The score is used only for relative ranking among agents at a given turn to determine who should speak next. This design enables decentralized speaker selection without introducing an additional controller model or handcrafted interruption rules, while keeping the mechanism lightweight and reproducible. At the same time, the present study does not claim that self-reported urgency is perfectly calibrated across tasks or models. Its calibration should therefore be interpreted as a limitation of the current framework and as an important direction for future work.

2) Speaker selection for the next turn: The next speaker is determined on the basis of the action plans of all agents. Candidates are limited to the agents that selected `speak` or `interrupt` as `Action`, and the agent with the highest `urgency` is selected as the next speaker. If multiple agents tie for the highest urgency, one of them is chosen at random so as to avoid bias from a fixed ordering.

If no candidate exists, the next turn becomes a silent turn.

Algorithm 1: Action Plan Prompt

```
You are {name}. You are debating to arrive at the
    correct answer to the question. Based on the
    debate information, generate a response to the
    instructions.

## Question
{topic}
# Debate Context
## Debate rules
- The debate ends when the shared public token
    budget of {token_budget} tokens is exhausted.
    Only revealed text counts toward this budget.
- Remaining public tokens available to all agents:
    {tokens_left} / {token_budget}.
- Only one member can speak per turn. The next
    speaker is selected from the highest urgency
    level.
- Each turn, one chunk at a time from the speaker's
    generated statement is revealed to all members.
    The current speaker may have more statement
    prepared and not yet revealed.

## Initial answers before debate begins (subject to
    change)
{initial_answer}

## Debate history
{turn_log}

## Your memory
{previous_thoughts}

## Event of this turn
{last_event}

## State
- This is turn {turn}. Use the remaining public
    tokens efficiently to reach the correct answer.

# Actions
You can take the following actions:
- 'listen': Use this when you listen to the
    speaker's argument until the end to deepen your
    understanding before speaking.
- 'interrupt': Use this when you interrupt the
    current speaker to begin speaking.

# Instructions
1. Based on the debate so far and the utterance of
    this turn, briefly explain your current
    internal thoughts such as your perspective on
    the responses to the questions, your action
    plan for the remaining turns, your reaction.
2. Determine the urgency for you to start talking
    now. Raise urgency only when 1) you can
    immediately correct a factual or logical error
    in the latest statement, 2) tokens_left is low
    and essential information must be shared soon,
    or 3) your current answer disagrees with the
    apparent majority in the debate history.
    Otherwise keep urgency low and listen.
3. Refer to the provided information and your
    current thought, decide the next turn's action
    you should take as {name}.
- While considering the possibility that someone
    may be mid-sentence, decide whether to
    interrupt and respond immediately to this
    turn's statement or listen to its completion.
4. Select the purpose of the action you have chosen.
5. Based on the debate information, output the
    answer to the question that you currently
    support the most.
- If the anticipated continuation of statement may
    resolve your concern, choose listen.
```

In a silent turn, no utterance is produced, and all agents update their action plans for the following turn.

3) Utterance generation and sentence-by-sentence disclosure: The agent selected as the next speaker generates an utterance on the basis of the debate history and its own action plan. The generated utterance is segmented into sentences and stored in an internal queue. Sentence segmentation is performed using the SpaCy natural language processing library*. We adopt sentences rather than fixed-length token chunks because sentences preserve semantic units more naturally and are closer to the minimal unit of human conversational contribution. During the debate, only one sentence is disclosed from the queue in each turn. This design makes it easier for other agents to intervene in the immediately following turn, even when an incorrect premise appears at the beginning of an utterance.

The disclosed sentences are accumulated as debate history and are then used as input for subsequent action-plan generation and utterance generation.

Sentence segmentation was adopted because sentences preserve semantic units more naturally than fixed-size token chunks and provide a practical unit for potential interruption. In this study, segmentation is performed deterministically with SpaCy for English debate outputs. However, segmentation quality may affect interruption timing, especially when generated text contains ellipsis, coordination, or non-canonical punctuation. The reported results should therefore be interpreted as evidence for the proposed mechanism under this segmentation setting rather than as a claim of segmentation invariance.

4) Definition of interruption: We define an interruption as a speaker change that occurs while a current speaker still has undisclosed sentences remaining in its queue and another agent is selected as the next speaker. When an interruption occurs, the next speaker begins speaking immediately in direct response to the preceding utterance. At that point, the original speaker's undisclosed sentences are discarded. By allowing this type of interruption, the proposed framework makes it easier for correction to occur before an explanation based on an incorrect premise expands over multiple sentences.

D. Final Answer Extraction

At each turn, the number of tokens in the disclosed utterance is counted incrementally and recorded as the cumulative number of public utterance tokens. The debate ends when this cumulative count reaches a predefined threshold. At the turn in which the threshold is reached, the `Answer` fields contained in the agents' action plans are collected as final-answer candidates, and the majority answer is adopted as the final answer.

IV. Experimental Settings

As the number of agents and the length of the discussion increase in multi-agent debate, more reasoning and rebuttal become possible, but the computational cost and total token consumption also grow rapidly, making it difficult to secure a large experimental scale and to compare conditions fairly. If debates are allowed to continue indefinitely, evaluation

*https://spacy.io/

becomes difficult not only because of computational constraints but also because the amount of discussion can differ substantially across conditions. For this reason, existing studies commonly evaluate under settings with upper bounds on the number of utterances or the number of agents [1], [14]. Similarly, prior work on dynamic turn-taking [2], [12] imposes explicit limits on game progression or discussion length in order to make evaluation feasible.

Based on these considerations, we set the upper limit of the total public utterance token budget to 500, which allowed sufficient room for discussion to develop while keeping execution costs manageable.

Table III summarizes the debate settings used in this study.

TABLE III. DEBATE SETTINGS USED IN THE EXPERIMENTS

Item	Value
Total Public Utterance Token Budget	500 tokens
Number of agents	3

The experiments in this study are designed as a controlled mechanism study rather than as a broad generalization study. A single model and a single benchmark are used so that differences among conditions can be attributed as directly as possible to turn-taking control. Likewise, conditioned initial states are used to reproduce situations in which correct and incorrect answers coexist and to isolate how each framework responds to misinformation already present in the debate. The results therefore establish the effectiveness of the proposed design under the evaluated setting, while broader validation across models, domains, and agent counts is left for future work.

A. LLM

We used Meta's Llama-3.1-8B-Instruct quantized to 8-bit[†]. The model parameters used in the experiments are listed in Table IV.

TABLE IV. LLM PARAMETERS USED IN THE EXPERIMENTS

Model	Llama-3.1-8B-Instruct
Context length	15000 tokens
Maximum tokens	1024 tokens
Temperature	0.3

B. Tasks and Dataset

To evaluate the effectiveness of the proposed framework, we adopted Massive Multitask Language Understanding (MMLU), which consists of multiple-choice questions across diverse subjects centered mainly on high-school-level knowledge [15]. MMLU covers a wide range of domains, including STEM, the humanities, and the social sciences, and has been widely used as a benchmark for evaluating the general knowledge and reasoning abilities of language models [16], [17], [18].

To isolate the effect of debate control, all compared conditions started from the same initial answers.

[†]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

In addition, to examine whether the introduction of interruption and silence can suppress the propagation of misinformation, we conditioned the initial answers. Specifically, the discussion began from a state in which a particular agent already held misinformation, allowing us to evaluate how each condition was affected by misinformation propagation.

More concretely, for each question randomly sampled from MMLU, we generated answers five times with a single LLM and stored the answer and reason obtained in each run. We retained only those questions for which both correct and incorrect answers appeared at least twice among the five outputs. At debate start, three of the five stored (answer, reason) pairs were assigned to the three agents so as to realize the desired configurations (two incorrect and one correct, or one incorrect and two correct), and these were embedded into the prompt as the agents' initial answers. This procedure allowed us to align the initial state—that is, the initial answers and reasons of all agents—across conditions and evaluate only the difference in debate control.

C. Compared Conditions

Rather than reproducing individual prior methods exactly, we compared abstracted conditions that capture representative turn designs in the literature so that the effects of speaking order and disclosure granularity could be disentangled. Specifically, the fixed-order condition abstracts fixed-turn MAD [1], [10], whereas the dynamic-order condition abstracts methods with dynamic speaker selection [2], [12]. To isolate the effect of the proposed debate control with interruption and silence, we compared the following three conditions.

1) Fixed order: Speakers alternated in a predetermined order, and each utterance was disclosed as a full generated message.

2) Dynamic order: At each turn, non-speakers declared their desire to speak based on urgency, and the next speaker was selected accordingly. However, utterances were disclosed at the message level, and speaker changes did not occur until the current speaker had finished.

3) Proposed framework: In addition to urgency-based dynamic speaker selection, the framework introduced sentence-by-sentence disclosure, urgency-based interruption, and silent turns, as described in Section III.

D. Evaluation Metrics

In each condition, debates were conducted on 1,000 questions from the constructed dataset. As evaluation metrics, we used the intermediate answers output by each agent in its action plan at each turn, evaluated the change in accuracy over turns, and thereby assessed the susceptibility of each method to misinformation. We also evaluated the accuracy of the final answers at the end of the debate to determine whether the proposed framework could improve accuracy while suppressing the propagation of misinformation. For reproducibility, the implementation, major prompts, evaluation scripts, and execution settings are publicly available at https://github.com/KimuraAkikazu/dynamicdebate.

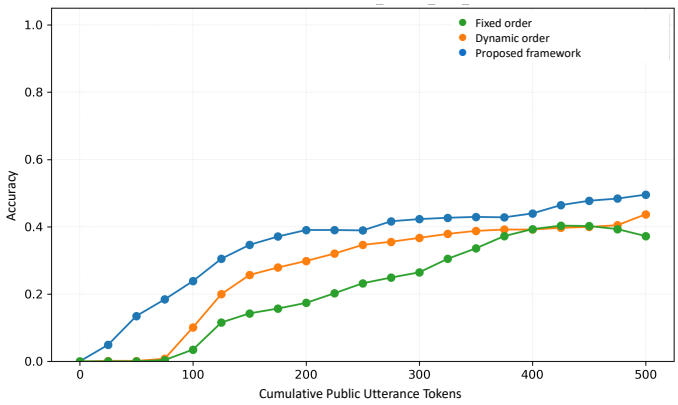


Fig. 1. Trajectory of intermediate-answer accuracy under each condition when two agents were initially incorrect.

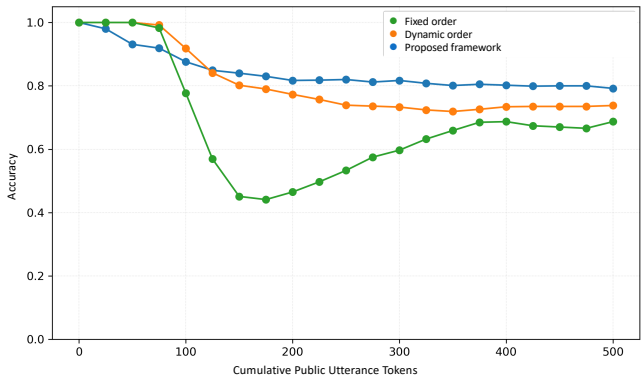


Fig. 2. Trajectory of intermediate-answer accuracy under each condition when one agent was initially incorrect.

V. Experimental Results

To examine the effectiveness of the proposed framework, we conducted evaluation experiments under the settings described in Section IV. In this section, we summarize the effects of the proposed framework from the viewpoints of final task accuracy, the trajectory of accuracy during the discussion, and answer changes caused by interruptions.

A. Accuracy

This subsection evaluates how the presence of agents holding misinformation affected debate accuracy. We compared final accuracy and the trajectory of intermediate answers under two settings that differed in the number of initially incorrect agents.

When two agents were initially incorrect, the proposed framework achieved 49.5%, which was 12.3 points higher than Fixed order (37.2%) and 5.8 points higher than Dynamic order (43.7%), as shown in Fig. 1 and Table V. Compared with Fixed order, the two dynamic conditions improved accuracy in the early stage of discussion and showed a more stable trajectory as the debate progressed. This early-stage gain was especially pronounced for the proposed framework.

TABLE V. Final Accuracy Under Each Condition

Condition	Accuracy
Fixed order	37.2%
Dynamic order	43.7%
Proposed framework	49.5%

These results indicate that the proposed framework is especially beneficial when the debate starts from an incorrect majority, suggesting that early interruption helps the correct side intervene before the incorrect majority becomes entrenched in the shared context.

When only one agent was initially incorrect, the proposed framework again achieved the highest accuracy at 79.2%, which was 10.5 points higher than Fixed order (68.7%) and 5.4 points higher than Dynamic order (73.8%), as shown in Fig. 2 and Table VI. In the two conditions without interruption, accuracy tended to drop sharply in the early stage. By contrast, in the proposed framework, the initial drop was comparatively suppressed and accuracy remained more stable as the debate progressed. In particular, under Fixed order, the incorrect agent was scheduled to speak first, causing a substantial initial drop in accuracy. Although some recovery was observed in the later stage, the final accuracy remained the lowest among the three conditions.

TABLE VI. Final Accuracy Under Each Condition When One Agent was Initially Incorrect

Condition	Accuracy
Fixed order	68.7%
Dynamic order	73.8%
Proposed framework	79.2%

In the one-incorrect-two-correct setting, the main advantage of the proposed framework is not only correction of the initially incorrect agent but also preservation of the correct agents from being pulled toward the wrong answer in the early stage of the debate.

B. Trajectory of Intermediate Answers by Agent

To examine how other agents' utterances influenced the answers of each agent, we collected the trajectories of intermediate answers for each method. In the legends of the graphs, Alex, Chris, and Jenny denote the three agents in the debate. The agent(s) holding the correct initial answer were Jenny in the two-incorrect-one-correct setting and Chris and Jenny in the one-incorrect-two-correct setting. In the Dynamic order condition, the next speaker was selected based on urgency, whereas in the Fixed order condition speaking rights alternated in the order Alex, Chris, and Jenny.

Fig. 3–5 show the trajectories of agent-wise intermediate-answer accuracy in the two-incorrect-one-correct setting. Under Fixed order (Fig. 3), the accuracy of the initially incorrect

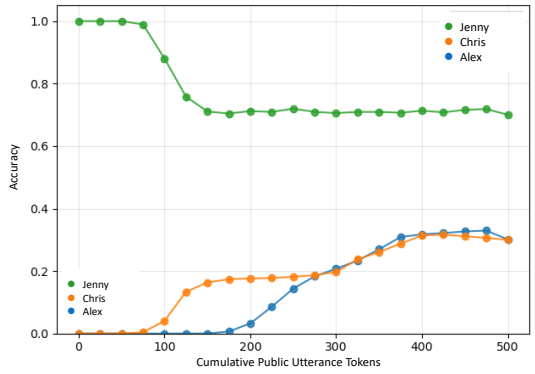


Fig. 3. Trajectory of intermediate-answer accuracy for each agent under Fixed order when two agents were initially incorrect.

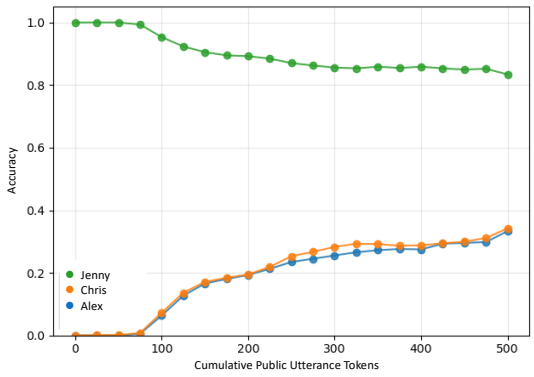


Fig. 4. Trajectory of intermediate-answer accuracy for each agent under Dynamic order when two agents were initially incorrect.

agents Alex and Chris remained close to 0 until the middle stage of the debate (around 200 tokens) and rose only to about 0.3 by the end. In contrast, the initially correct agent Jenny dropped sharply from 1.0 to around 0.7 in the early stage (around 100–150 tokens) and did not recover substantially thereafter. This suggests that the incorrect agents spoke early, formed the shared context, and made the correct agent more vulnerable to the influence of incorrect premises. Under Dynamic order (Fig. 4), Jenny remained around 0.85 while declining only gradually, indicating that the correct side was less likely to be dragged toward the incorrect side than under Fixed order. However, the improvement of Alex and Chris started only after about 100 tokens and plateaued around 0.3, making it difficult for the majority to flip from incorrect to correct. Under the proposed framework (Fig. 5), Jenny stayed at about the same high level as in Dynamic order, while Alex and Chris began improving earlier and reached about 0.4 by the end. The earlier correction of the incorrect side contributed to the higher majority-vote accuracy shown in Table V.

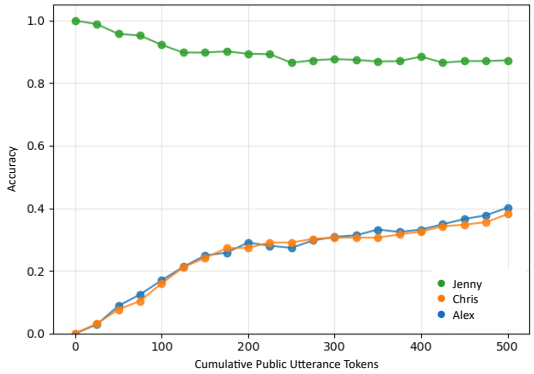


Fig. 5. Trajectory of intermediate-answer accuracy for each agent under the proposed framework when two agents were initially incorrect.

Fig. 6–8 show the corresponding trajectories in the one-incorrect-two-correct setting. Under Fixed order (Fig. 6), the early utterances of the incorrect agent Alex caused the accuracies of the correct agents Chris and Jenny to drop sharply from 1.0 to around 0.6 near 100–200 tokens. Although they later recovered, this early drop lowered the majority-vote accuracy (Table VI). Under Dynamic order (Fig. 7), the decline of Chris and Jenny was more gradual and converged around 0.8, avoiding the sharp deterioration observed under Fixed order. At the same time, Alex also improved from around 100 tokens onward and reached about 0.4 by the end. Under the proposed framework (Fig. 8), Chris and Jenny remained around 0.86–0.9 throughout the debate, while Alex improved step by step from the early stage and also reached about 0.4. In other words, the proposed framework maximized final accuracy by minimizing the correct side's being pulled toward incorrect answers while also accelerating the correction of the incorrect side.

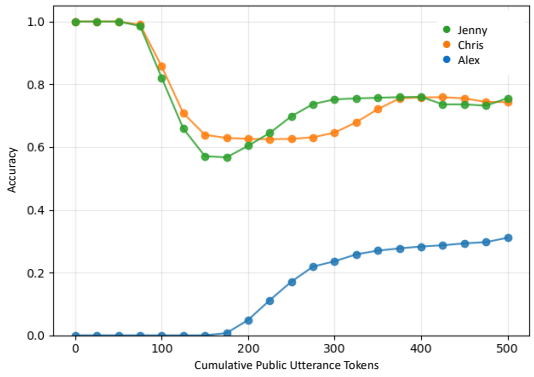


Fig. 6. Trajectory of intermediate-answer accuracy for each agent under Fixed order when one agent was initially incorrect.

VI. Analysis and Discussion

This section analyzes why the proposed framework improved accuracy from the perspective of the debate process, based on the results in Section V.

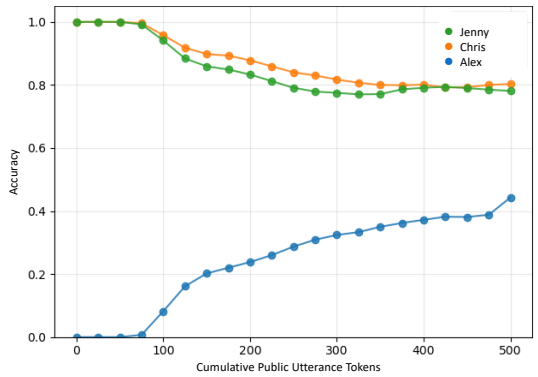


Fig. 7. Trajectory of intermediate-answer accuracy for each agent under Dynamic order when one agent was initially incorrect.

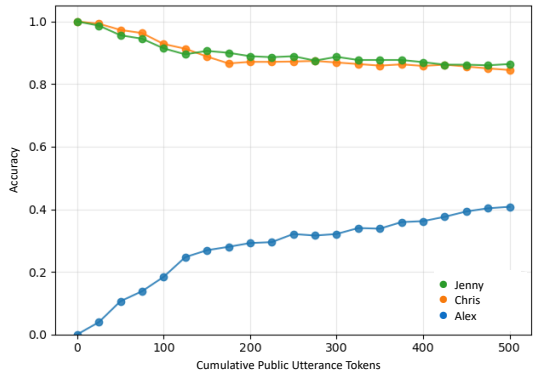


Fig. 8. Trajectory of intermediate-answer accuracy for each agent under the proposed framework when one agent was initially incorrect.

A. Improvement in Final Accuracy

As shown in Table V, in the two-incorrect-one-correct setting the proposed framework achieved a final accuracy of 49.5%, outperforming Fixed order (37.2%) and Dynamic order (43.7%). As shown in Table VI, the proposed framework also achieved 79.2% in the one-incorrect-two-correct setting, outperforming Fixed order (68.7%) and Dynamic order (73.8%).

The two-incorrect-one-correct setting starts with an incorrect majority. When incorrect premises are shared under such a setting, the majority error is likely to be reinforced during the debate. The fact that the proposed method improved accuracy in this setting suggests that it promoted correction on the incorrect side and guided the debate toward flipping the majority from incorrect to correct.

Fig. 1 shows that, under the proposed framework, majority-vote accuracy increased from the early stage of the debate and then remained at a comparatively high level. In contrast, in the one-incorrect-two-correct setting, Fixed order and Dynamic order both exhibited an early drop in accuracy, whereas the proposed framework suppressed that drop (Fig. 2). These results suggest that sentence-by-sentence disclosure and urgency-based interruption allowed intervention and repair to begin before an incorrect claim expanded into a long utterance, thereby suppressing the spread of misinformation while accelerating updates on the incorrect side.

B. Effects of Interruptions and Completion

To further explain why the proposed framework improved accuracy, this subsection compares the effects of 1) interruptions and 2) completion of an utterance on updates to the intermediate answers of the other agents. Here, *completion* refers to the case in which a speaker finishes disclosing the entire utterance (sentence queue) without being interrupted.

1) Evaluation units and scope: An interruption event is defined as the case in which one speaker still has undisclosed sentences remaining in the queue and another agent is selected as the next speaker (Section III). A completion event is defined as the case in which the speaker fully discloses the generated sentence queue and the utterance ends without interruption. In Fixed order and Dynamic order, interruptions do not occur, so every speaking turn is counted as a completion event. In the proposed framework, by contrast, interruptions can occur, so completion events are a subset of the speaking turns that finish without interruption.

For each event, we compared the `Answer` of each agent immediately before the event (just before the speaker change) with the `Answer` after the event (at the time the interrupter finished speaking, in the case of interruption, or when the utterance finished, in the case of completion). The resulting changes in the listeners' `Answer` values were classified into three categories. Because this subsection evaluates the effect of intervention when correct and incorrect answers coexist, cases in which all agents were correct or all agents were incorrect immediately before the event were excluded from the analysis.

The three categories are defined as follows, where s denotes the speaker of the event (either the interrupter or the speaker whose utterance completed) and L denotes the listeners (all agents other than s).

- **Improved**: The speaker s was on the correct side, and at least one listener in L changed from incorrect to correct.
- **Worsened**: The speaker s was on the incorrect side, and at least one listener in L changed from correct to incorrect.
- **Unchanged**: Neither of the above occurred.

For clarity, the percentages reported below are event-level proportions computed over all evaluated events under the stated condition. This event-level view complements final-answer accuracy by showing whether a turn type tends to repair or spread misinformation during the debate process.

2) Effects of interruptions: Under the proposed framework, interruptions yielded an improvement rate of 15.2% and a worsening rate of 7.1% in the two-incorrect-one-correct setting, so improvement exceeded deterioration (Table VII). In the one-incorrect-two-correct setting, the corresponding values were 12.2% and 5.2%, showing the same tendency (Table VIII). Because the worsening rate was lower than the improvement rate in both settings, interruptions can be interpreted as contributing, on average, to suppressing misinformation propagation and advancing repair earlier. The

percentages of improved, worsened, and unchanged reported in this subsection were calculated with all evaluated events as the denominator.

TABLE VII. EFFECTS OF INTERRUPTIONS IN THE PROPOSED FRAMEWORK (TWO INCORRECT, ONE CORRECT)

Metric	Value
Evaluated events	533
Improved	15.2%
Worsened	7.1%
Unchanged	77.7%

TABLE VIII. EFFECTS OF INTERRUPTIONS IN THE PROPOSED FRAMEWORK (ONE INCORRECT, TWO CORRECT)

Metric	Value
Evaluated events	1308
Improved	12.2%
Worsened	5.2%
Unchanged	82.6%

Taken together, Tables VII and VIII indicate that interruption functions more often as an early repair mechanism than as a source of instability under the evaluated setting.

3) Effects of completion: Next, we compared the updates observed immediately after completion events across conditions. In the two-incorrect-one-correct setting, the proposed framework achieved an improvement rate of 14.3% and a worsening rate of 6.5%, again indicating that improvement exceeded deterioration (Table IX). Dynamic order showed a higher improvement rate of 18.1%, but its worsening rate was also high at 11.3%, leaving a smaller net advantage. Under Fixed order, the worsening rate exceeded the improvement rate, resulting in a net negative effect. In the one-incorrect-two-correct setting as well, the proposed framework showed an improvement rate of 10.5% and a worsening rate of 3.7%, indicating a tendency to accumulate improvements while suppressing deterioration (Table X). In the proposed framework, only speaking turns that finished without interruption were counted as completion events.

TABLE IX. COMPARISON OF COMPLETION EFFECTS ACROSS CONDITIONS (COMPLETION WITHOUT INTERRUPTION) IN THE TWO-INCORRECT-ONE-CORRECT SETTING

Condition	Improved	Worsened
Proposed framework	14.3%	6.5%
Dynamic order	18.1%	11.3%
Fixed order	13.6%	18.5%

TABLE X. COMPARISON OF COMPLETION EFFECTS ACROSS CONDITIONS (COMPLETION WITHOUT INTERRUPTION) IN THE ONE-INCORRECT-TWO-CORRECT SETTING

Condition	Improved	Worsened
Proposed framework	10.5%	3.7%
Dynamic order	17.6%	9.2%
Fixed order	13.6%	13.0%

These completion results suggest that the proposed framework does not merely increase the frequency of intervention. Rather, it maintains a more favorable balance between improvement and deterioration even when an utterance is allowed to finish, which is consistent with more stable information sharing than the other two conditions.

4) When interruptions are effective: To analyze at what points in the debate interruptions were effective, we computed the success rates of improvement and deterioration as a function of the cumulative number of public utterance tokens at the time the interruption occurred. Because interruption effects were more clearly visible in the two-incorrect-one-correct setting, we report this analysis for that setting.

Specifically, we binned the cumulative number of public utterance tokens at the time of interruption in increments of 50 tokens and computed the following two conditional probabilities for each bin. Here, *improve_success_rate* is the proportion of interruptions that led to improvement, conditioned on the interrupter being on the correct side, whereas *worsen_success_rate* is the proportion that led to deterioration, conditioned on the interrupter being on the incorrect side.

$$\text{improve_success_rate} = \frac{I_{succ}}{I_{succ} + I_{fail}} \tag{1}$$

$$\text{worsen_success_rate} = \frac{W_{succ}}{W_{succ} + W_{fail}} \tag{2}$$

Here, I_{succ} denotes the number of cases in which, after an interruption by a correct-side speaker, at least one listener changed from incorrect to correct, and I_{fail} denotes the number of cases in which no such update occurred. Similarly, W_{succ} denotes the number of cases in which, after an interruption by an incorrect-side speaker, at least one listener changed from correct to incorrect, and W_{fail} denotes the number of cases in which no such update occurred.

As shown in Fig. 9, the improve_success_rate was relatively high in the early stage, around 50–199 tokens, and exceeded the worsen_success_rate in many bins. This can be interpreted as indicating that, while the shared context is still short, an interruption can initiate repair immediately after an incorrect premise appears and is therefore more likely to suppress the following spread of misinformation. By contrast, around 300–349 tokens, the worsen_success_rate exceeded the improve_success_rate, suggesting that in the later stage of the debate, interruptions made when few public utterance tokens remain may destabilize consensus formation. This tendency indicates that the effectiveness of interruptions should be controlled according to the stage of the discussion.

C. Limitations

This study has several limitations. First, the evaluation uses a single base model (Llama-3.1-8B-Instruct) and a single benchmark (MMLU), so the results should not be interpreted as evidence of universal effectiveness across models or domains. Second, the experiments fix the number of agents to three and constrain discussion by a 500-token shared public budget; different agent counts or budget regimes may yield different interruption dynamics. Third, urgency is produced by the LLM itself and used as a relative coordination signal, but its calibration is not externally validated in the present study. Fourth, sentence segmentation is performed with SpaCy, and

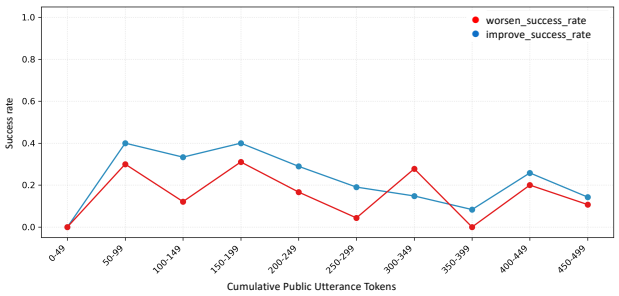


Fig. 9. Improvement and deterioration success rates of interruptions in the proposed framework (50-token bins, two incorrect and one correct).

the robustness of the framework to segmentation errors or alternative segmentation policies is not evaluated. Finally, the initial states are intentionally conditioned to include both correct and incorrect answers so that misinformation propagation can be examined under controlled conditions. This design improves internal comparability across conditions, but it does not fully represent all naturally occurring debate setups.

D. Future Work

Several directions follow from these limitations. First, broader validation is needed across additional LLMs, domain-specific benchmarks, and tasks whose reasoning unfolds differently from multiple-choice question answering. Second, the urgency signal should be calibrated more explicitly, for example by comparing self-reported urgency with externally detected factual conflicts, confidence estimates, or agreement patterns among agents. Third, the framework should be evaluated under larger agent populations and different public-token budgets to clarify how interruption frequency, silence, and consensus stability scale with group size. Fourth, alternative disclosure units and segmentation strategies, including clause-level, discourse-based, or adaptive chunking schemes, should be compared to determine when sentence-level disclosure is most effective. Fifth, because the framework updates action plans at fine granularity, future work should quantify trade-offs among accuracy, generated tokens, number of model calls, and execution time.

In addition, previous studies have reported that assigning personality traits to agents may improve answer accuracy in collaborative tasks and diversify debate [19], [11]. Future work should examine how overall debate accuracy changes when such personality assignments are combined with the interruption- and silence-aware turn-taking control proposed in this study. In particular, it will be important to analyze how differences in personality traits affect the frequency of interruptions and silence, the timing of error correction, and the tendency for unnecessary interruptions to occur. Such analysis may also help realize flexible dialogue in which opinions are presented at effective moments through interruption and silence when the framework is extended to human–LLM discussions.

VII. Conclusion

This study presented an interruptible multi-agent debate framework that combines sentence-level disclosure, urgency-based speaker selection, interruption, and silent turns under a shared public-token budget. The central goal was to reduce the spread of incorrect premises by allowing listeners to intervene before a misleading message expands into a long explanation. Under the evaluated setting, the proposed framework achieved the highest final accuracy among the compared conditions in both the two-incorrect-one-correct and one-incorrect-two-correct settings. Analysis of intermediate answers and event-level updates further suggested that interruption more often supported correction than deterioration. These findings indicate that finer-grained turn control can improve debate stability when correct and incorrect answers coexist. At the same time, the present results are bounded by the evaluated model, benchmark, agent count, and segmentation setting. Broader validation and improved urgency calibration remain important next steps.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP23K28377, JP24H00714, JP25K15109, JP25K03190, JP25K03232, JP22K12157 and The Telecommunications Advancement Foundation.

References

[1] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," *arXiv preprint arXiv:2305.14325*, 2023.

[2] S. Bailis, J. Friedhoff, and F. Chen, "Werewolf arena: A case study in LLM evaluation via social deduction," *arXiv preprint arXiv:2407.13943*, 2024.

[3] Y. Feng, Y. Wang, S. Cui, B. Faltings, M. Lee, and J. Zhou, "Unraveling misinformation propagation in LLM reasoning," in *Findings of the Association for Computational Linguistics: EMNLP 2025*. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 11 683–11 707. [Online]. Available: https://aclanthology.org/2025.findings-emnlp.627/

[4] A. Amayuelas, X. Yang, A. Antoniades, W. Hua, L. Pan, and W. Y. Wang, "Multiagent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 6929–6948.

[5] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[6] E. A. Schegloff, G. Jefferson, and H. Sacks, "The preference for self-correction in the organization of repair in conversation," *Language*, vol. 53, no. 2, pp. 361–382, 1977.

[7] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: Communicative agents for "mind" exploration of large language model society," *arXiv preprint arXiv:2303.17760*, 2023.

[8] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu *et al.*, "Autogen: Enabling next-gen LLM applications via multi-agent conversation," *arXiv preprint arXiv:2308.08155*, 2023.

[9] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun, "Chatdev: Communicative agents for software development," *arXiv preprint arXiv:2307.07924*, 2023.

[10] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, "Chateval: Towards better LLM-based evaluators through multi-agent debate," in *International Conference on Learning Representations*, 2024.

[11] J. Zhang, X. Xu, N. Zhang, R. Liu, B. Hooi, and S. Deng, "Exploring collaboration mechanisms for LLM agents: A social psychology view," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024, pp. 14 544–14 607.

[12] R. Nonomura and H. Mori, "Who speaks next? multi-party AI discussion leveraging the systematics of turn-taking in murder mystery games," *Frontiers in Artificial Intelligence*, vol. 8, p. 1582287, 2025.

[13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 24 824–24 837.

[14] Y. Zeng, W. Huang, L. Jiang, T. Liu, X. Jin, C. T. Tiana *et al.*, "S2-mad: Breaking the token barrier to enhance multi-agent debate efficiency," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 9393–9408.

[15] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2021.

[16] A. Chowdhery *et al.*, "PaLM: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[17] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[18] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[19] A. Kimura, K. Fukuda, R. Orihara, Y. Tahara, and Y. Sei, "The impact of personality trait integration into LLM agents on collaborative tasks," in *Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence*, 2025, 3J6-GS-5-04.

Secure User Authentication Model Using Identity-Based Encryption (IBE) Scheme: Challenges, Techniques, and Trends

Raja Farah Sharima Raja Muhamad Danial[1], Nazhatul Hafizah Kamarudin[2], Abdul Ghafar Jaafar[3]
Centre for Cyber Security-Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia [1, 2]
Faculty of Artificial Intelligence, Universiti Teknologi Malaysia (UTM), 54100 Kuala Lumpur, Malaysia [3]

***Abstract*—User authentication is the crucial component in ensuring that digital identities are securely authenticated when performing sensitive digital transactions. Building resilience against emerging cyber threats like identity theft and Man-in-the-Middle (MITM) attacks remains an ongoing challenge. Identity-Based Encryption (IBE) is a cryptographic method in the category of public key cryptography that allows the use of unique identifiers, such as email addresses, as public keys for encryption, making key management easier than traditional public key cryptography. This study presents a comprehensive overview of secure authentication techniques using IBE schemes proposed by researchers, emphasizing techniques that have been explored and introduced. The review examines current IBE schemes, including blockchain-based techniques, and analyzes the security features, application domains, implementation requirements, and challenges. By incorporating findings from previous works, this study identifies common challenges that impede real-world implementation and suggests a potential approach for combining alternative security methods to improve authentication robustness. The purpose of this work is to provide researchers and practitioners with a comprehensive review of secure authentication models, as well as practical insights to assist in developing an implementable authentication solution for real-world secure digital transactions.**

***Keywords*—User authentication; Identity-Based Encryption (IBE); digital identity; network security**

I. Introduction

The rapid growth of digital services has made secure user authentication a fundamental and crucial requirement for any digital transaction provided by service providers, whether private or Government agencies. Many domains, such as finance, healthcare and the Internet of Things (IoT), have increasingly relied on digital identities to enable access to sensitive services and data. Consequently, the authentication process has become a critical security control, as failures in identity verification can result in severe consequences such as identity theft, unauthorized access, privacy breaches, and financial losses.

Identity authentication methods have gone through evolution since they were introduced, that is, Single-Factor Authentication (SFA), Two-Factor Authentication (2FA) and Multi-Factor Authentication (MFA) [1]. All these authentication methods can be combined with additional factors to further strengthen the security level of user identity information, such as security tokens, user biometric information, cryptographic or encryption methods, secure environment, and many others.

An identity-based cryptographic method, called Identity-Based Encryption (IBE), was introduced by [2]. It is an extension of the Public Key Encryption (PKE) method, where the user's public key is some special information about the user, for example, the user's email address. This means that a sender who has access to the system's public parameters can use the text value of the recipient's name or email address as a key to encrypt the message they want to send. This approach simplifies key management and reduces the need for Public Key Infrastructure (PKI).

However, the IBE method can only provide examples of identity-based signatures. Hence, in order to address the problems encountered, various improved IBE methods have been introduced by researchers from time to time. In 2001, Dan Boneh and Matthew K. Franklin helped to solve the IBE issue by introducing a pair-based cryptography named the Boneh–Franklin scheme [3] and [4] has introduced an encryption method based on the quadratic residue, that is, the message is encrypted one bit at a time, but suffers from inefficiency due to the high expansion of the ciphertext. The IBE method by [5] uses bilinear coupling, which provides a different approach to IBE.

Although the IBE method has been improved, it faces a limited usage issue due to the absolute trust that is essential to the Private Key Generator (PKG) [6] and [7]. The PKG can generate any user's private key and potentially decrypt (or sign) messages without permission, and this is a weakness of the IBE method [8]. Therefore, various alternative systems and continuous studies are needed to reduce the level of trust required in PKG [9]. Certificate-based encryption, secure key production cryptography, and certificateless cryptography [10] are among the alternative systems that have been developed to release the PKG component.

The evolution of the IBE has reflected significant advances in cryptographic practice aimed at improving security, efficiency, and usability. Accordingly, ongoing research is expected to help strengthen the level of security and reliability of cryptographic systems against emerging cyber threats. Thus, this study aims to conduct a comprehensive analysis of the IBE schemes currently in use that are efficient and effective. By

knowing the most practical and secure IBE schemes, this study will provide potential possible combinations of technologies in further enhancing the user identification and securing user's personal information security and privacy in performing digital services. This study also aims to provide future research on developing an ideal and general model of user authentication that can be adopted by any field, with the ability to mitigate possible security risk and be one of the factors to increase users' confidence and trust to use digital services.

A. Motivation

The main purpose of this study is to examine the current method used for the user identity authentication process using the IBE scheme, as well as emerging technologies that is suitable for adaptation to increase the level of security and privacy of the user's personal data. The primary motivation of this study is as follows:

- Examine the IBE scheme that is effective and secure for user identity authentication.
- Assess the advantages, limitations, or challenges of all IBE schemes currently used.
- Identify opportunities for enhancement and innovation in user identity authentication.
- Minimize cybersecurity risks related to user identity authentication, such as Man-in-the-Middle (MITM), identity theft, and phishing.
- Gain knowledge on issues related to user identity authentication.

B. Contributions

There are several contributions from this review study that may help other researchers on user authentication methods and cybersecurity issues by addressing significant barriers, synthesizing knowledge, and investigating potential future works or innovations related to IBE.

Besides, this study will provide useful direction for researchers, programmers, as well as organizations intending to deploy or enhance the security level of user authentication methods in their applications.

The contributions include the following:

- Provide a comprehensive analysis of the current landscape of user authentication methods, highlighting their key application fields, key advantages and critical security challenges.
- Discusses significant challenges in user authentication techniques and proposes feasible and relevant perspectives for improvement.
- Provide significant insights on user authentication model applicable across various fields.
- Provide ongoing efforts to further improve the resilience and robustness of user authentication methods by highlighting areas that require further research especially related to security issues.

C. Research Questions

In achieving the objectives of this study, four (4) research questions have been identified with its rationale and are detailed as follows:

- RQ1: What fields of study use the IBE scheme?
- RQ2: What is the problem statement of the previous research IBE scheme?
- RQ3: What are the limitations or weaknesses faced in using the IBE scheme?
- RQ4: What are the new technologies or trends in the IBE authentication field?

The research questions are determined to explore the current user authentication model using the IBE scheme. This study will investigate and discuss all questions outlined in detail.

D. Research Objectives

Based on the research questions, information on common themes of fields, problems, limitations and new technologies will be identified. The research questions are also used to achieve the following research objectives:

- RO1: to explore application fields which used IBE as an authentication method done by previous researchers.
- RO2: to understand problems addressed or issues faced by previous researchers on the IBE authentication method.
- RO3: to recognize limitations or weaknesses faced by the current IBE scheme proposed or tested.
- RO4: to explore new technologies or models that can further enhance the scheme.

E. Paper Organization

The organization of this study is designed to address all the research questions and research objectives outlined.

The study structures are as follows:

- Section II presents and analyzes related previous works.
- Section III details the research methodology used, including the search strategy and the inclusion and exclusion criteria.
- Section IV reviews the implementation of IBE scheme in various fields.
- Section V presents issues and challenges faced in IBE scheme.
- Section VI details out the limitations in IBE scheme.
- Section VII provides information on future perspectives and enhancements.
- Section VIII discusses issues related to IBE scheme and ways to overcome.
- Section IX concludes and summarizes the analysis and findings of the study based on the research objectives.

List of abbreviations used in this study is listed in Table I below:

TABLE I. LIST OF ABBREVIATIONS

Acronym	Definition
2FA	Two-Factor Authentication
BC-IBE	Backward Compatible Identity-Based Encryption
CA	Central Authority
DKER	Decryption Key Exposure Resistance
DRIBE	Delegated Revocable Identity-Based Encryption
GDPR	General Data Protection Regulation
IBE	Identity-Based Encryption
IoT	Internet of Things
KGC	Key Generation Center
LWE	Learning with Errors
MFA	Multi-Factor Authentication
MITM	Man-in-the-Middle
MPTCP	Multi-Path TCP
PKE	Public Key Encryption
PKG	Private Key Generator
PKI	Public Key Infrastructure
RIBE	Revocable Identity-Based Encryption
RHIBE	Revocable Hierarchical Identity-Based Encryption
SFA	Single-Factor Authentication

II. RELATED WORKS

Numerous relevant studies have been carried out to examine, analyze and evaluate different user authentication models. A study on the trust issue faced in Identity-Based Encryption conducted by [8] has highlighted the need for a more secure and deployable privacy solution by reducing the probability of privacy breaches to an arbitrarily small value. Based on a study conducted by [11] on numerous IBE schemes, the main problem that exists in the IBE scheme, which is escrow, has been identified, and suggested that dependency on PKG needs to be overcome to address the issue of trust in third parties, since it will indirectly affect the privacy of communication between users.

The Revocable Hierarchical Identity-Based Encryption (RHIBE) scheme introduced and discussed by [12] focuses on adaptively secure construction methods to address the weaknesses in previous methods that can only achieve selective security, rather than adaptive security. Similarly, since revocation functionality is crucial in cryptographic systems and existing RIBE schemes require long public parameters or composite-order groups, [13] has presented study on an efficient Revocable Identity-Based Encryption (RIBE) scheme. Furthermore, the need for a secure Identity-Based Encryption scheme with perfect forward secrecy for network security protocols to address classic IBE which lacks the perfect forward secrecy have been discussed by [14].

In protecting the privacy of data in Internet of Things (IoT) applications, [15] has proposed an effective and useful IBE method with a Revocable IBE (RIBE) feature where the Private Key Generator (PKG) can cancel the user's account if the private key has been exposed. While issues on compatibility requirements due to existing systems will lose access to previous ciphertexts after the key update process has been discussed in detail by [16]. The study in [17]'s paper has addressed the efficient key revocation in RIBE schemes by exploring delegating update key generation to a cloud server due to inefficiencies with the update key sizes issue faced in previous schemes.

Additionally, a study conducted by [18] has discussed the issue of privacy in smart contracts and the challenges of implementing smart contracts that preserve privacy, because it was found that transaction details are openly disclosed in traditional smart contracts. Therefore, researchers introduce a smart contract privacy protection mechanism and develop a new blockchain model to preserve smart contract privacy.

As a result of the study carried out by [19], there are various challenges related to digital identity management in the context of Web 3.0, especially in the need to obtain a more private, secure and manageable Internet experience. Issues of difficulty in ensuring interoperability between various platforms and privacy rules in a decentralized environment are also discussed.

In addition, researchers also found that the most important and main problem is the transition from a traditional identity system to a self-sovereign identity system that allows users to manage the sharing of their personal information independently without involving a Central Authority (CA) in increasing the level of user control and security [19].

Comparison of the authentication scheme features is detailed in Table II. A summary of previous works is described in Table III. Fig. 1 illustrates the Identity-Based Encryption (IBE) architecture.

IBE provides more secure and enhanced security features compared to traditional authentication models. But IBE has its own limitations and advantages as well. The main limitations of IBE are inherent weaknesses and issues such as reliance on PKG which requires centralized trust, and if compromised, it will lead to a single point of failure issue. In addition, the key revocation difficulty issue in the basic IBE scheme also causes compromised key revocation to affect the implementation of IBE in dynamic user environments. Although IBE has some limitations, it has many advantages and has been widely used in many fields such as banking, IoT and healthcare.

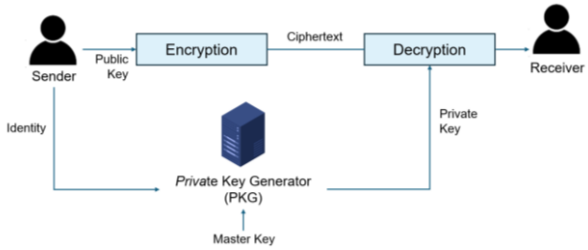


Fig. 1. IBE architecture.

TABLE II. COMPARISON OF AUTHENTICATION SCHEMES

Authentication Scheme	Security Focus	Threats Addressed	Technologies Integrated	Real-World Applicability
Single-Factor Authentication (SFA)	Basic identity verification using a single factor (knowledge, possession or inherence).	Weak passwords, phishing, credential stuffing, brute-force attacks.	Passwords, PINs, swipe cards, biometric fingerprints or facial recognition.	Personal or computer devices (password of the smartphone, computer, laptop), email logins.
Two-Factor Authentication (2FA)	Requires two different authentication factors to strengthen security. Typically using one knowledge factor and one possession or inherence factor.	Phishing, password guessing, credential stuffing.	Combination of password and OTP (One Time Password via SMS / mobile applications), hardware tokens, combination of password and biometric information.	Online banking, email, corporate VPNs, social media accounts.
Multi-Factor Authentication (MFA)	Requires two or more different authentication factors to create multiple layers of defense and enhanced security level.	Stolen credentials, phishing, brute force, Man-in-the-Middle (MITM) attacks, replay attacks, account takeovers.	Passwords, biometrics information (fingerprint, facial recognition), hardware keys (USB/NFC), push notifications.	Industry standard to secure sensitive accounts in fields of healthcare, IoT, government systems, financial services, enterprise environments.
Identity-Based Encryption (IBE)	Cryptographic authentication method using identity-linked encryption keys rather than static passwords.	Impersonation, unauthorized data access, interception, Man-in-the-Middle (MITM) attacks via cryptographic means.	Public Key Infrastructure (PKI) (public keys are derived from user identities), secure key management, encryption algorithms.	Secure messaging, email encryption, IoT environments and applications requiring encrypted communications linked to identity.

TABLE III. SUMMARY OF PREVIOUS RESEARCH

Paper	Issue Discussed	Contribution
[8]	Overcoming trust issue in IBE scheme	Review a privacy scheme without non-collusion assumptions to allow users to reduce private key exposure probability by employing multiple intermediate CAs for enhanced privacy and the scheme separates the PKG from the intermediate CAs.
[11]	Comprehensive study and analysis on numerous IBE schemes	Provides detailed information on numerous IBE schemes with its advantages, disadvantages and potential attacks.
[12]	Adaptive security of IBE scheme	Provides proof on adaptive security under simple static assumptions and addresses challenges of dual system encryption in Revocable Hierarchical Identity-Based Encryption (RHIBE) through security proof.
[13]	Revocation requirements in IBE scheme	Proposes a Revocable Identity-Based Encryption (RIBE) scheme that achieves adaptive security, Decryption Key Exposure Resistance (DKER), and constant-size public parameters over prime-order bilinear groups by employing the Seo-Emura technique, which combines specific IBE schemes to enhance security and efficiency.
[14]	Importance of perfect forward secrecy	Provide details of Learning with Errors (LWE) problem based on Quantum Identity-Based Encryption (QIBE) scheme and its advantages.
[15]	Importance of secure communication in Internet of Things (IoT) applications	Proposes an efficient RIBE scheme for IoT applications with its detail algorithms and utilizes ISO/IEC 18033-5 SM9 encryption for security.
[16]	Compatibility requirements in IBE scheme	Introduction of Backward Compatible Identity-Based Encryption (BC-IBE) to solve backward compatibility problems in IBE systems and supports revocation of expired private keys.
[17]	Exploring delegation of update key generation	Introduces the concept of RIBE that can be delegated (Delegated Revocable Identity-Based Encryption (DRIBE)) where the process of delegation of update key generation to the cloud server will be implemented.
[18]	Preserving privacy in smart contracts	Introduce privacy-protection mechanisms for smart contracts and develop a new blockchain model for privacy-preserving smart contracts. Details comparison of computation and storage overheads with existing schemes are also provided.
[19]	Challenges related to digital identity management	Provide information on digital identity evolution and how users or organizations use digital identities to control their online appearance inside the Web 3.0 architecture.
This paper	User authentication using IBE scheme in secure and privacy-protected environment	This paper provides comprehensive reviews on various IBE schemes across different fields and functions. Findings from trustworthy databases are combined and classified, offering distinct points of view based on deployment fields, issues, challenges and integration of evolving technologies in proposing future and further research areas. This will aid in identifying important gaps and offer a roadmap for the enhanced and advancement of user authentication methods that are deployable in various fields.

The advantage of IBE is that it does not require distribution of public keys, as they will be derived from known identities hence a separate PKI is not required. This will simplify and increase the efficiency of key management and reduce overhead. IBE is also flexible as it uses any unique string such as email or username as the public key which allows embedding additional information in the identity to dynamically control the validity of the key. And it is also scalable as the entire system uses a single master key pair maintained by the PKG which makes key generation and distribution more efficient for many users.

In the banking sector, IBE provides secure and scalable financial transactions where it can protect sensitive banking transactions by directly encrypting data with user identities (such as account numbers or emails) without the use of complex PKI, thus reducing the risk of MITM attacks. Additionally, the combination of IBE with digital ID authentication can enhance user authentication and data privacy by securely binding identity data to a cryptographic key.

For the Internet of Things (IoT) sector, IBE provides simplification of cryptographic key management in resource-

constrained devices by eliminating cumbersome PKI setup and it also supports scalable and flexible key distribution in dynamic IoT networks. Device ID with embedded identity also provides easy authentication and encryption. In terms of communication and storage, it supports secure cloud communication and storage suitable for IoT environments.

While for the healthcare sector, IBE allows secure exchange of confidential information associated with patient or provider identities without the need for complex PKIs and simplifies key management for multiple healthcare systems and practitioners. Besides, it also ensures secure cloud data transmission and medical record storage.

In conclusion, IBE provides more cost-effective security management compared to traditional PKI by reducing management overhead and making enterprise-level encryption scalable and affordable. IBE also provides flexible encryption that is directly linked to user identities and simplified key management. This is highly beneficial for sectors like banking, IoT and healthcare that require secure and scalable communications. However, the security risks associated with key escrow and reliance on a single trusted authority need to be addressed properly.

III. Research Methodology

The methods and strategies applied in this study are explained in this section. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol [20] is the methodology adopted for this work.

A. Search Strategy

The established digital library databases that have been identified to be used for obtaining related articles are Web of Science, IEEE Xplore, Scopus and Wiley. The publication papers are the recent papers, from 2020 to 2025. Keywords that were used in the searching process are "user authentication" or "user authentication model" or "Identity Based Encryption (IBE)" or "IBE".

These keywords were searched either in the title or abstract using Boolean operators like "AND" and "OR" together with parenthesis. The key string used in the searching process is: ("User Authentication" AND "IBE"), ("User Authentication Model" AND "IBE"), ("User Authentication" AND ("Identity-Based Encryption (IBE)" OR "Identity-Based Encryption") and ("User Authentication Model" AND ("Identity-Based Encryption (IBE)" OR "Identity-Based Encryption")).

In total, 414 papers were discovered from these four databases and after assessing the papers, 124 duplicates or unrelated papers were eliminated.

B. Inclusion and Exclusion Criteria for Selection

Details of the inclusion and exclusion criteria are described in Table IV.

TABLE IV. Inclusion and Exclusion Criteria

Inclusion Criteria	
IC1	Papers that were published between 2020 and 2025
IC2	Papers that are accessible and downloadable
IC3	Paper related on user authentication models using IBE scheme
IC4	Papers that provide design, scheme, or implementation of IBE in the user authentication model
IC5	Papers that are using English language
Exclusion Criteria	
EC1	Papers that are not related to the work and research question or objectives of this work
EC2	Papers that are not accessible as full text
EC3	Papers discussing user authentication method that are other than IBE scheme
EC4	Papers which studies do not include design, scheme, or implementation of IBE in the user authentication model
EC5	Papers using other than English language.

The selection and filtering process using the PRISMA methodology is illustrated in Fig. 2.

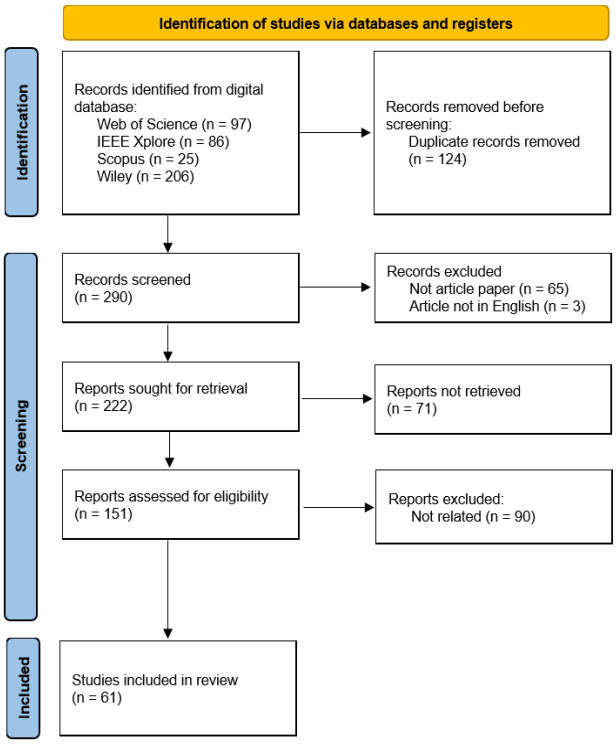


Fig. 2. PRISMA methodology flow diagram.

Based on the PRISMA statement, 61 papers have been identified and will be thoroughly reviewed and scrutinized in order to determine the current state of the user authentication method.

IV. Implementation of Identity-Based Encryption (IBE) in Various Fields

Based on the papers reviewed, various fields of studies have been identified. This section categorizes fields of study into three main fields: Security and Information Sharing, Internet of Things (IoT), and Network and Infrastructure.

Details of each field, which consist of the number of papers, equivalent representation in terms of percentage and presentation of an extensive overview of current state for every field identified, are shown in Fig. 3 and Table V.

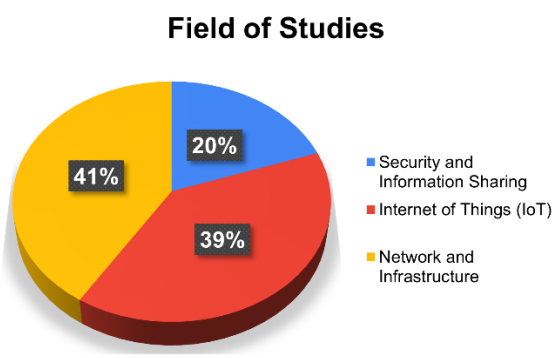


Fig. 3. Field of studies.

TABLE V. CATEGORIES OF FIELDS DISCOVERED IN THE PREVIEWED PAPERS.

Field of Studies	Papers
Security and Information Sharing	[6], [8], [11], [12], [13], [14], [16], [18], [19], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35]
Internet of Things (IoT)	[15], [36], [37], [38], [39], [40], [41], [42], [43]
Network and Infrastructure	[7], [9], [10], [17], [24], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63]

A. Security and Information Sharing

Security is the most critical factor, especially in the field of information and technology. In line with security principles, confidentiality, integrity and availability, users' personal information must always be ensured secured at the most optimal level. Based on research papers that have been examined, 20% of them have highlighted that security issues have to be addressed accordingly in ensuring users' information and privacy are well preserved [8], [19], [18].

Likewise, a study by [27] has discussed challenges in ensuring outsourced data integrity, while [32] highlighted challenges related to outsourcing in cloud storage systems. A blockchain model for privacy-preserving smart contracts has also been proposed by [18] as a measure to increase the level of information security and user privacy.

In sectors related to security and information sharing, many enterprises have adopted IBE scheme to streamline corporate communications and secure email by using user identities such as email addresses directly as public keys, removing the need for revocation infrastructure, certificate authorities and certificates.

B. Internet of Things (IoT)

Currently, IoT has been widely used in many fields and industries. Thus, user authentication is vital in ensuring only authorized users can access facilities provided accordingly. 39% of the reviewed papers which are in IoT field highlight that having an efficient IBE scheme are crucial [15], [30], [38].

Similarly, [30] also highlighted the need to enhance user privacy in cryptographic schemes. While [39] have proposed solution to encounter issues related to data and privacy which was compromised in cloud storage environment. In addition, [41] has suggested that in order to handle identity revocation, it is crucial to provide data confidentiality and source authentication. Furthermore, [15] emphasized the importance of having secure communications in IoT applications through user authentication.

Implementation of IBE in IoT, such as Lightweight Authentication in IoT devices, is widely deployed in many sectors due to the reduced complexity of key distribution and management, as device identities (such as the serial number or device name) can directly serve as public keys. This will reduce the overhead in IoT contexts, which often have network and trust infrastructure limitations as well as restricted computational resources.

Additionally, the use of Hierarchical Identity-Based Encryption (HIBE) in IoT applications that have multiple levels of devices and access hierarchies (such as gateways, edge devices and sensors) is very useful due to the hierarchical key representation feature provided by HIBE. This hierarchical structure will also facilitate effective access control management in IoT ecosystems where devices may be organized under multiple administrative domains.

C. Network and Infrastructure

The highest fields from the reviewed papers are from the network and infrastructure field, which contributes 41%. According to [55], attacks such as session hijacking, traffic diversion, and eavesdropping during the initial handshake are possible. Therefore, security is essential for the successful implementation of MPTCP (Multi-Path TCP).

The Signal Protocol and similar cryptographic systems face several security challenges, including vulnerabilities to MITM attacks and users cannot fully trust service providers, as there are possibilities of malicious servers to distribute forged public keys [24].

Based on a study conducted by [50], secure data transfer is the challenges in fog computing and IoT ecosystems. Therefore, researchers propose IBE scheme called Hierarchical Identity-Based Architecture for Fog Computing (HIBAF) to enhance data security by providing performance enhancement in data processing and management compared to traditional cloud environments.

The difficulty of assuring data security, integrity and reliability in cloud storage systems have been examined by [58]. Likewise, [17] have conducted study related to delegation of update key generation to a cloud server by introducing the Delegated RIBE (DRIBE) concept.

Mobile Ad Hoc Networks (MANETs) are an example of a variant of IBE that has been prototyped in the networking and infrastructure-related sectors. By using device IDs such as phone numbers for encryption and authentication, they can enable effective, certificate-free secure communication without the need for a traditional PKI. In addition, IBE also has been utilized for network access control, simplifying user and device authentication without depending on certificate management systems by encoding identities in IBE private keys.

V. ISSUES AND CHALLENGES IN IBE

In accordance with the identified and evaluated papers, main issues that have been discovered related to IBE scheme in user

authentication model have been listed in this review. These issues are discovered and classified into two main categories, based on the problem statements and issues highlighted in the reviewed papers, as shown in Fig. 4 and Table VI.

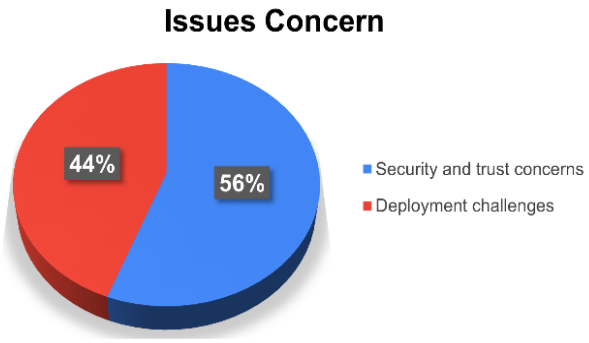


Fig. 4. Categories of issues of concern.

TABLE VI. ISSUES DISCUSSED IN THE PREVIEWED PAPERS

Issues	Papers
Security and trust concerns	[6], [9], [10], [11], [12], [14], [15], [16], [18], [22], [23], [24], [25], [26], [27], [30], [31], [33], [34], [38], [39], [41], [42], [43], [48], [49], [50], [51], [52], [53], [56], [61], [62]
Deployment challenges	[7], [8], [13], [15], [17], [19], [21], [24], [28], [29], [32], [35], [36], [37], [39], [40], [44], [45], [46], [47], [54], [55], [57], [58], [59], [60], [63]

A. Security and Trust Concerns

As shown in Fig. 4, the most critical and difficult issue to overcome is related to the level of users' personal information security with regard to verifying users' identities digitally and users' trust in the technology and online services offered. This is driven by various factors and is evident when the number of reported cybercrime cases is increasing despite the implementation of additional security features. This, in turn, pushes towards a higher level of improvement required in strengthening the level of security of users' personal information in order to increase and maintain the level of user trust.

Current password-based methods expose users to security risks thus encouraging a wider implementation of IBE in real-world contexts [22]. In IBE schemes, the issue of escrow [24], [9], [31], [11] and trust on third parties [51] is the primary and common issue. Thus, security and privacy concerns towards user's information need to be addressed [56] by enhancing security features like mutual authentication and key revocation [24].

Studies on security solutions to protect against various threats are highly needed in maintaining user's trust [52] and ensuring the confidentiality in communication amidst quantum threats [38]. Besides, [18] have identified that smart contract implementation faced challenges in preserving privacy. Other than that, most schemes only achieve selective security, not adaptive security [12], [31].

B. Deployment Challenges

The second issue of concern identified from the reviewed papers are the implementation challenges. Challenges related to the implementation of the IBE scheme in the real-world environment are the complexity of the model, interoperability across different platforms, practicality, cost, and computing time involved, level of effectiveness, reliability, flexibility, and the need for specialized cryptographic expertise.

Some of the IBE schemes are very complex and difficult to achieve tight security [21], [57], [63] and face complexity related to certificate [39], [32] and key [37] management. Additionally, [19] have highlighted the difficulties in ensuring interoperability among various platforms and the implications of privacy regulations, such as the right to be forgotten in a decentralized environment. And there are inefficiency issues related to keys, such as the update key sizes [17] and key regeneration for non-revoked users [13].

VI. LIMITATIONS

As indicated in the reviewed papers, limitations of the IBE scheme that have been discovered from the previous work are analyzed in this review. The limitations are categorized into three main categories, which are usability in real-world environments, vulnerabilities and the need for expertise. Details of all the categories are provided in Fig. 5 and Table VII.

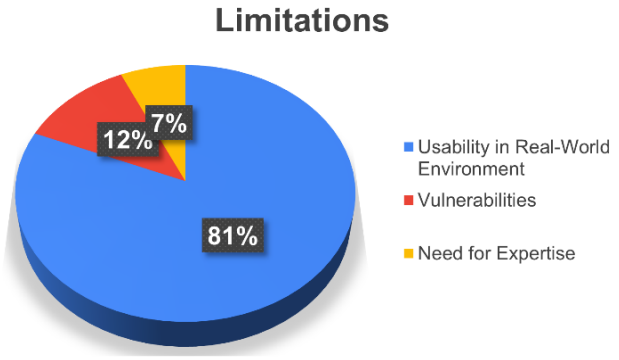


Fig. 5. Categories of limitations.

TABLE VII. LIMITATIONS IN THE PREVIEWED PAPERS

Issues	Papers
Usability in Real-World Environment	[6], [7], [8], [9], [10], [11], [12], [13], [15], [16], [17], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [36], [37], [39], [41], [42], [43], [45], [46], [47], [48], [49], [50], [51], [52], [54], [55], [56], [57], [58], [60], [61], [62], [63]
Vulnerabilities	[14], [22], [24], [35], [40], [44], [53]
Need for Expertise	[18], [19], [21], [59]

A. Usability in Real-World Environment

Proposals or ideas for a solution or a theoretically proposed model are very useful and helpful, by proving it through formal or theoretical methods will make it more meaningful and adaptable in the real-world environment. The key escrow issue would hinder the deployment of any IBE scheme since the Key Generation Center (KGC) have the ability to decrypt all ciphertext which would lead to potential security risks [23]. Thus, implementation of IBE in real-world environments are still limited [6] due to certain factors which need to be addressed before wider deployment is feasible.

B. Vulnerabilities

Overcoming possible threats and vulnerabilities with regard to crucial information, such as a user's digital identity and their personal information in online services are vital. Example of a situation, during user registration process, the Signal Protocol is vulnerable to Man-in-the-Middle (MITM) attacks due to malicious server actions [24]. Vulnerabilities to attacks, which might be due to the usage of low-entropy passwords, will become a challenge [53] even though the scheme used already deploys necessary security measures. And as discussed by [35], theoretical explanations of security might not completely encompass real-world vulnerabilities.

C. Need for Expertise

The emergence of new technology is very valuable in the development of today's technological world. But to adopt any technology, it requires more than just merely utilizing the technology. A comprehensive and detailed plan, together with relevant knowledge of the particular technology and requirements, is compulsory to drive the technology adoption and achieve the expected success. Therefore, expertise and skilled developers are required to ensure the successfulness for any solution's implementation.

In paper [18], it does recognize the need to have a developer with deep cryptography knowledge, especially in handling complexity in the blockchain environment. Similarly, [19] also raised about the difficulty of understanding and adopting underlying technologies such as blockchain and decentralized networks, as well as becoming one of the challenges to create a robust digital identity ecosystem.

VII. Future Perspectives and Enhancements

In advancing and securing the user authentication process by adopting relevant emerging technologies, more detailed research needs to be conducted. By exploring, identifying, analyzing, and choosing the most reliable and efficient technology would bring advantages to many fields and industries, especially those that require secure user authentication.

Since the authentication process is the most commonly used by people and organizations, it must be equipped with robust security measures. Therefore, any proposed model or solution must be tested in a real-world environment using relevant security tools and proved to be secure. This study would provide insights on possible and potential areas to be further examined by researchers. Future enhancements categories are described in Table VIII and Fig. 6.

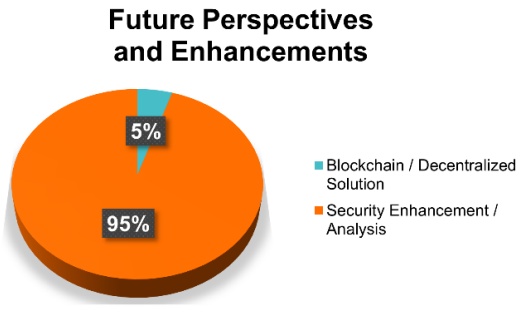


Fig. 6. Categories of future perspectives and enhancements.

TABLE VIII. Future Perspectives and Enhancements in the Previewed Papers.

Issues	Papers
Blockchain / Decentralized Solution	[19], [22], [61]
Security Enhancement / Analysis	[6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [21], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [62], [63]

A. Blockchain / Decentralized Solution

Blockchain is one of the new technologies that has been explored by many industries. Many research related to user authentication also has considered and recognized the potential usage of blockchain technology by adopting it in their solution. According to [61], the studies have suggested future work on comprehensive rights divisions in blockchain-empowered metaverse applications, as existing editable blockchains cannot support user preferences on readability governance, which violates the GDPR rights to restrict processing.

Based on studies and solutions by [22], future works suggested are on decentralized storage in enhancing the solution's management beyond authentication. This will contribute in providing a more efficient and reliable authentication model in terms of performance, security, and ease of use.

B. Security Enhancement / Analysis

By deploying additional security features, it will not only be able to gain and maintain user trust towards the online services provided, but it will also help in establishing an organization's reputation. Many biological attributes have been explored and deployed as reliable and trustable security features that are hardly compromised. Similarly, a study by [50] has proposed a future plan in combining their proposed scheme with face or iris biometrics for enhanced protection. While [35] aimed to develop a biometric identity-based blind signature scheme.

VIII. Discussion

The security of information, especially personal information and the user's identity, must be protected and preserved from any cyber threat and ensure that it cannot be accessed, used, disclosed or modified without permission by unauthorized parties or cyber criminals who are increasingly worried by the public, especially users of digital services. The emergence of various digital applications which require user identity verification such as face recognition through the selfie process and identity card images as part of a verification and registration process to obtain digital services offered is one of the risks of personal information being exposed to cyber threats. This is due to the security level of the application is difficult to be guaranteed and proved as completely secure, and the location of the information's storage and who has access to the data are also unknown.

Various identity verification methods have been introduced and there are methods that can be combined with additional factors to further strengthen the level of security of user identity

information, such as security tokens, user biometric information, cryptographic or encryption methods, secure environment, and many others. However, the most secure and efficient identity verification model in terms of computing or execution duration, computational cost, and proven to be safe through security tests is difficult to be found [17], [15], [23].

In addition, for the issue of key regeneration for non-revoked users, which was found to be inefficient by [13], a method based on prime-order bilinear groups that includes server-assisted secure methods and selected ciphertexts could be used. While to protect data privacy in Internet of Things (IoT) applications, the Revocable IBE (RIBE) feature, which was proposed by [15], can be adopted as it was found to be effective.

As discussed by [18], although issues related to privacy in smart contracts could be overcome by deploying a new blockchain model to preserve the smart contract privacy as the protection mechanism, a security model for selected ciphertext attacks should be established, as well as security for various types of cryptographic assumptions should be explored.

By referring to study that have been done by [19], where numerous challenges were encountered related to digital identity management, further research to create a more robust digital identity ecosystem in order to achieve a more secure environment and ensure the seamless integration across different platforms is highly needed.

The main limitation of IBE is its reliance on PKGs, which are a single point of failure and a valuable target for attackers. This can be overcome by using Artificial Intelligence (AI) or Machine Learning (ML) to improve anomaly detection, as AI/ML can monitor PKG activity for unusual patterns and identify attempted compromises or unauthorized access attempts more quickly and effectively. Additionally, AI can analyze the global threat landscape and predict potential attack vectors targeting PKGs.

For the issue of the inefficiency of IBE for practical use due to bitwise message encryption, which leads to ciphertext expansion and high computational cost, especially for decryption, ML can be used to analyze the computational cost of IBE algorithms, identify bottlenecks, and suggest optimizations to speed up processes such as decryption without compromising security.

Thus, more detailed studies still need to be carried out as an effort to strengthen user identity verification methods that include efficient cryptography or encryption methods, secure authentication environment from any threat, interoperability across platform, as well as proof of performance and security levels that need to be obtained through security performance tests using security equipment and real-world environment scenarios in having a more appropriate and practical models that can be used in various fields.

IX. Conclusion

Based on the review and analysis from previous works that have been carried out by all the researchers in various topics that are related to user authentication using IBE scheme, numerous schemes and models can be adopted to ensure that the user identity verification process, especially user's information and privacy of data, is protected from any possible risks such as identity theft and cyber-attack like Man-in-the- Middle (MITM) attack. All the proposed solutions have its own advantages and disadvantages, either from the aspect of computing cost for the algorithm used or time taken to execute the algorithm, and the level of deployability in a real-world environment. This will depend on the needs and requirements, which may vary in different fields and industries. However, there are solutions that can be categorized as ideal and suitable to be a sample model in developing a more general model that can be deployed in various fields or industries.

In addition, as proposed and suggested by most of the researchers, integrating the existing IBE scheme with other relevant technologies can increase the level of security and overcome the inherent issues of privacy and trust faced in the IBE scheme. And there are many studies that have deployed new technologies such as blockchain in their solutions [18], [10], [33]. Thus, a general, lightweight, and practical user identity verification model using an IBE scheme that integrates with more secure technologies or environments such as blockchain or other emerging technologies, is expected to be able to overcome the limitations and challenges faced in existing IBE schemes and ensure wide range of implementation across industries, especially in public sectors.

In conclusion, this study aims to provide insights to address challenges faced in existing IBE schemes and proposed areas of study on potential solutions as suggested that may contribute to the development of user authentication models. As digital metaverse keeps evolving, all possible security risks must be taken into serious consideration, especially regarding privacy, accountability and data integrity in order to gain and increase user trust. Important factors such as performance, reliability, efficiency, flexibility and scalability also must be considered in the studies. This study also aims to contribute significant insights related to IBE schemes by analyzing challenges and opportunities and provide a comprehensive understanding of the current state and future directions in order to enhance the security level of user authentication in our daily needs and businesses as user digital identity are crucial in digital services.

Acknowledgment

The authors would like to express sincere gratitude to the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, for the support in conducting this study. This study was supported by the Universiti Kebangsaan Malaysia (UKM) under the Research Grant Scheme under Grant FRGS/1/2025/ICT07/UKM/02/1.

References

[1] A. Ometov, S. Bezzateev, N. Mäkitalo, S. Andreev, T. Mikkonen, and Y. Koucheryavy, "Multi-factor authentication: A survey," Cryptography, vol. 2, no. 1, pp. 1–31, Mar. 2018, doi: 10.3390/cryptography2010001.

[2] A. Shamir, "Identity-Based Cryptosystems and Signature Schemes," in LNCS, 1985, pp. 47–53. doi: 10.1007/3-540-39568-7_5.

[3] D. Boneh and M. Franklin, "Identity-Based Encryption from the Weil Pairing," SIAM Journal on Computing, vol. 32, no. 3, pp. 586–615, Jan. 2003, doi: 10.1137/S0097539701398521.

[4] C. Cocks, "An Identity Based Encryption Scheme Based on Quadratic Residues," 2001, pp. 360–363. doi: 10.1007/3-540-45325-3_32.

[5] M. Kasahara, "Cryptosystems based on pairing," in The 2000 Symposium on Cryptography and Information Security, 2000, pp. 26–28.

[6] C. Adams, "Security Analysis of a Privacy-Preserving Identity-Based Encryption Architecture," Journal of Information Security, vol. 13, no. 04, pp. 323–336, 2022, doi: 10.4236/jis.2022.134018.

[7] T. L. Astrizi and R. Custódio, "Seamless Transition to Post-Quantum TLS 1.3: A Hybrid Approach Using Identity-Based Encryption," Sensors, vol. 24, no. 22, p. 7300, Nov. 2024, doi: 10.3390/s24227300.

[8] C. Adams, "Improving User Privacy in Identity-Based Encryption Environments," Cryptography, vol. 6, no. 4, 2022, doi: 10.3390/cryptography6040055.

[9] Y. Liang, G. Di Crescenzo, H. Wang, and Z. Patni, "Efficient Identity-Based Encryption with Minimal Server Trust," in 2024 43rd International Symposium on Reliable Distributed Systems (SRDS), IEEE, Sep. 2024, pp. 104–114. doi: 10.1109/SRDS64841.2024.00020.

[10] R. Saha et al., "A Blockchain Framework in Post-Quantum Decentralization," IEEE Trans Serv Comput, vol. 16, pp. 1–12, 2023, doi: 10.1109/TSC.2021.3116896.

[11] A. Karrothu and J. Norman, "A systematic analysis of Identity Based Encryption (IBE)," International Journal of Knowledge-Based and Intelligent Engineering Systems, vol. 25, no. 3, pp. 343–356, 2021, doi: 10.3233/KES-210078.

[12] K. Lee, "Revocable hierarchical identity-based encryption with adaptive security," Theor Comput Sci, vol. 880, pp. 37–68, Aug. 2021, doi: 10.1016/j.tcs.2021.05.034.

[13] K. Emura, J. H. Seo, and Y. Watanabe, "Efficient revocable identity-based encryption with short public parameters," Theor Comput Sci, vol. 863, pp. 127–155, Apr. 2021, doi: 10.1016/j.tcs.2021.02.024.

[14] W. Gao, L. Yang, D. Zhang, and X. Liu, "Quantum Identity-Based Encryption from the Learning with Errors Problem," Cryptography, vol. 6, no. 1, p. 9, Feb. 2022, doi: 10.3390/cryptography6010009.

[15] Y. Sun, P. Chatterjee, Y. Chen, and Y. Zhang, "Efficient Identity-Based Encryption with Revocation for Data Privacy in Internet of Things," IEEE Internet Things J, vol. 9, no. 4, pp. 2734–2743, Feb. 2022, doi: 10.1109/JIOT.2021.3109655.

[16] J. Kim, "Backward Compatible Identity-Based Encryption," Sensors, vol. 23, no. 9, p. 4181, Apr. 2023, doi: 10.3390/s23094181.

[17] K. Lee, "Delegate and Verify the Update Keys of Revocable Identity-Based Encryption," IEEE Access, pp. 1–1, 2023, doi: 10.1109/ACCESS.2023.3280253.

[18] H. Yin, Y. Zhu, G. Guo, and W. C.-C. Chu, "Privacy-Preserving Smart Contracts for Confidential Transactions Using Dual-Mode Broadcast Encryption," IEEE Trans Reliab, vol. 73, no. 2, pp. 1090–1103, Jun. 2024, doi: 10.1109/TR.2023.3328146.

[19] S. L. Nita and M. I. Mihailescu, "A Novel Authentication Scheme Based on Verifiable Credentials Using Digital Identity in the Context of Web 3.0," Electronics (Basel), vol. 13, no. 6, p. 1137, Mar. 2024, doi: 10.3390/electronics13061137.

[20] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," Syst Rev, vol. 10, no. 1, p. 89, Dec. 2021, doi: 10.1186/s13643-021-01626-4.

[21] R. Langrehr and J. Pan, "Tightly Secure Hierarchical Identity-Based Encryption," Journal of Cryptology, vol. 33, no. 4, pp. 1787–1821, Oct. 2020, doi: 10.1007/s00145-020-09356-x.

[22] P. Szalachowski, "Password-Authenticated Decentralized Identities," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 4801–4810, 2021, doi: 10.1109/TIFS.2021.3116429.

[23] K. Emura, S. Katsumata, and Y. Watanabe, "Identity-based encryption with security against the KGC: A formal model and its instantiations," Theor Comput Sci, vol. 900, pp. 97–119, Jan. 2022, doi: 10.1016/j.tcs.2021.11.021.

[24] S. Liu, Y. Shao, H. Luo, and H. Di, "IBE-Signal: Reshaping Signal into a MITM-Attack-Resistant Protocol," Security and Communication Networks, vol. 2022, pp. 1–24, Jul. 2022, doi: 10.1155/2022/8653453.

[25] M. Clear and H. Tewari, "Anonymous Homomorphic IBE with Application to Anonymous Aggregation," Cryptography, vol. 7, no. 2, p. 22, Apr. 2023, doi: 10.3390/cryptography7020022.

[26] B. Xu et al., "T-FIM: Transparency in Federated Identity Management for Decentralized Trust and Forensics Investigation," Electronics (Basel), vol. 12, no. 17, p. 3591, Aug. 2023, doi: 10.3390/electronics12173591.

[27] J. Sun, G. Xu, T. Zhang, X. Yang, M. Alazab, and R. H. Deng, "Privacy-Aware and Security-Enhanced Efficient Matchmaking Encryption," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 4345–4360, 2023, doi: 10.1109/TIFS.2023.3294725.

[28] D. Hofheinz, J. Koch, and C. Striecks, "Identity-Based Encryption with (Almost) Tight Security in the Multi-instance, Multi-ciphertext Setting," Journal of Cryptology, vol. 37, no. 2, p. 12, Apr. 2024, doi: 10.1007/s00145-024-09496-4.

[29] B. Zuo, J. Li, Y. Zhang, and J. Shen, "Identity-Based Online/Offline Encryption Scheme from LWE," Information, vol. 15, no. 9, p. 539, Sep. 2024, doi: 10.3390/info15090539.

[30] S. Dong, Z. Zhao, B. Wang, W. Gao, and S. Zhang, "SM9 Identity-Based Encryption with Designated-Position Fuzzy Equality Test," Electronics (Basel), vol. 13, no. 7, p. 1256, Mar. 2024, doi: 10.3390/electronics13071256.

[31] K. Asano, K. Emura, and A. Takayasu, "More Efficient Adaptively Secure Lattice-Based IBE with Equality Test in the Standard Model," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E107.A, no. 3, Mar. 2024, doi: 10.1587/transfun.2023CIP0021.

[32] J. Du, G. Dong, J. Ning, Z. Xu, and R. Yang, "Identity-based controlled delegated outsourcing data integrity auditing scheme," Sci Rep, vol. 14, no. 1, p. 7582, Mar. 2024, doi: 10.1038/s41598-024-58325-y.

[33] D. Marchsreiter, "Towards quantum-safe blockchain: Exploration of PQC and public-key recovery on embedded systems," IET Blockchain, vol. 5, no. 1, Jan. 2025, doi: 10.1049/blc2.12094.

[34] H. Wang, K. Chen, Q. Xie, and Q. Meng, "Post-Quantum Secure Identity-Based Matchmaking Encryption," IEEE Trans Dependable Secure Comput, vol. 22, no. 1, pp. 833–844, Jan. 2025, doi: 10.1109/TDSC.2024.3418984.

[35] X. Shan, L. You, and G. Hu, "Two Efficient Constructions for Biometric-Based Signature in Identity-Based Setting Using Bilinear Pairings," IEEE Access, vol. 9, pp. 25973–25983, 2021, doi: 10.1109/ACCESS.2021.3057064.

[36] K. Emura, A. Takayasu, and Y. Watanabe, "Efficient identity-based encryption with Hierarchical key-insulation from HIBE," Des Codes Cryptogr, vol. 89, no. 10, pp. 2397–2431, Oct. 2021, doi: 10.1007/s10623-021-00926-z.

[37] D. Unal, A. Al-Ali, F. O. Catak, and M. Hammoudeh, "A secure and efficient Internet of Things cloud encryption scheme with forensics investigation compatibility based on identity-based encryption," Future Generation Computer Systems, vol. 125, pp. 433–445, Dec. 2021, doi: 10.1016/j.future.2021.06.050.

[38] D. Dharminder, A. K. Das, S. Saha, B. Bera, and A. V. Vasilakos, "Post-Quantum Secure Identity-Based Encryption Scheme using Random Integer Lattices for IoT-enabled AI Applications," Security and Communication Networks, vol. 2022, pp. 1–14, Jul. 2022, doi: 10.1155/2022/5498058.

[39] M. Zhao and Y. Ding, "Dual-Server Identity-Based Encryption with Authorized Equality Test for IoT Data in Clouds," Security and Communication Networks, vol. 2022, pp. 1–12, Oct. 2022, doi: 10.1155/2022/4905763.

[40] Y. Su, Y. Li, and Z. Cao, "Gait-Based Privacy Protection for Smart Wearable Devices," IEEE Internet Things J, vol. 11, no. 2, pp. 3497–3509, Jan. 2024, doi: 10.1109/JIOT.2023.3296650.

[41] X. Hu, L. Wang, L. Gu, and Y. Ning, "A Bilateral Access Control Data Sharing Scheme for Internet of Vehicles," IEEE Internet Things J, vol. 11, no. 22, pp. 36748–36762, Nov. 2024, doi: 10.1109/JIOT.2024.3420176.

[42] N. Gao et al., "Ciphertext Retrieval With Identity Bidirectional Authentication and Matrix Index in IoT," IEEE Internet Things J, vol. 11, no. 1, pp. 889–903, Jan. 2024, doi: 10.1109/JIOT.2023.3288131.

[43] H. Xiong et al., "Heterogeneous Signcryption with Equality Test for IIoT Environment," IEEE Internet Things J, vol. 8, no. 21, pp. 16142–16152, Nov. 2021, doi: 10.1109/JIOT.2020.3008955.

[44] Y. Zhang, Y. Liu, Y. Guo, S. Zheng, and L. Wang, "Adaptively Secure Efficient (H)IBE over Ideal Lattice with Short Parameters," Entropy, vol. 22, no. 11, p. 1247, Nov. 2020, doi: 10.3390/e22111247.

[45] T. A. Lone et al., "Securing communication by attribute-based authentication in HetNet used for medical applications," EURASIP J Wirel Commun Netw, vol. 2020, no. 1, p. 146, Dec. 2020, doi: 10.1186/s13638-020-01759-5.

[46] H. Ji, H. Zhang, L. Shao, D. He, and M. Luo, "An efficient attribute-based encryption scheme based on SM9 encryption algorithm for dispatching and control cloud," Conn Sci, vol. 33, no. 4, pp. 1094–1115, Oct. 2021, doi: 10.1080/09540091.2020.1858757.

[47] A. Wu, D. Feng, M. Zhang, A. Yang, and J. Chi, "Privacy-Preserving Bilateral Multi-Receiver Matching with Revocability for Mobile Social Networks," IEEE Trans Mob Comput, vol. 23, no. 12, pp. 11080–11090, Dec. 2024, doi: 10.1109/TMC.2024.3390036.

[48] S. Liu, L. Chen, G. Wu, H. Wang, and H. Yu, "Blockchain-Backed Searchable Proxy Signcryption for Cloud Personal Health Records," IEEE Trans Serv Comput, vol. 16, no. 5, pp. 3210–3223, Sep. 2023, doi: 10.1109/TSC.2023.3272770.

[49] W. Shen, J. Yu, M. Yang, and J. Hu, "Efficient Identity-Based Data Integrity Auditing with Key-Exposure Resistance for Cloud Storage," IEEE Trans Dependable Secure Comput, vol. 20, no. 6, pp. 4593–4606, Nov. 2023, doi: 10.1109/TDSC.2022.3228699.

[50] M. Whaiduzzaman et al., "HIBAF: A data security scheme for fog computing," Journal of High Speed Networks, vol. 27, no. 4, pp. 381–402, Nov. 2021, doi: 10.3233/JHS-210673.

[51] I. Gomaa, E. Abd-Elrahman, A. Hamdy, and E. M. Saad, "Automated Security Assessment for IDaaS Framework," Wirel Pers Commun, vol. 116, no. 4, pp. 3465–3490, Feb. 2021, doi: 10.1007/s11277-020-07860-8.

[52] C. Jiang, C. Huang, Q. Huang, and J. Shi, "A Multi-Source Big Data Security System of Power Monitoring Network Based on Adaptive Combined Public Key Algorithm," Symmetry (Basel), vol. 13, no. 9, p. 1718, Sep. 2021, doi: 10.3390/sym13091718.

[53] A. Karati, R. Amin, S. H. Islam, and K.-K. R. Choo, "Provably Secure and Lightweight Identity-Based Authenticated Data Sharing Protocol for Cyber-Physical Cloud Environment," IEEE Transactions on Cloud Computing, vol. 9, no. 1, pp. 318–330, Jan. 2021, doi: 10.1109/TCC.2018.2834405.

[54] A. Takayasu and Y. Watanabe, "Revocable identity-based encryption with bounded decryption key exposure resistance: Lattice-based construction and more," Theor Comput Sci, vol. 849, pp. 64–98, Jan. 2021, doi: 10.1016/j.tcs.2020.10.010.

[55] A. S. Almuflih, K. Popat, V. V. Kapdia, M. R. N. M. Qureshi, N. Almakayeel, and R. E. Al Mamlook, "Efficient Key Exchange Using Identity-Based Encryption in Multipath TCP Environment," Applied Sciences, vol. 12, no. 15, p. 7575, Jul. 2022, doi: 10.3390/app12157575.

[56] Z. Zhou, B. B. Gupta, A. Gaurav, Y. Li, M. D. Lytras, and N. Nedjah, "An Efficient and Secure Identity-Based Signature System for Underwater Green Transport System," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 9, pp. 16161–16169, Sep. 2022, doi: 10.1109/TITS.2022.3148166.

[57] K. Emura and A. Takayasu, "A Generic Construction of CCA-Secure Identity-Based Encryption with Equality Test against Insider Attacks," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E106.A, no. 3, p. 2022CIP0012, Mar. 2023, doi: 10.1587/transfun.2022CIP0012.

[58] X. Zhang, C. Huang, D. Gu, J. Zhang, and H. Wang, "BIB-MKS: Post-Quantum Secure Biometric Identity-Based Multi-Keyword Search over Encrypted Data in Cloud Storage Systems," IEEE Trans Serv Comput, pp. 1–1, 2021, doi: 10.1109/TSC.2021.3112779.

[59] B. B. Gupta, A. Gaurav, C.-H. Hsu, and B. Jiao, "Identity-Based Authentication Mechanism for Secure Information Sharing in the Maritime Transport System," IEEE Transactions on Intelligent Transportation Systems, pp. 1–9, 2021, doi: 10.1109/TITS.2021.3125402.

[60] R. Wang et al., "Securing Topology Control in SDWSNs Using Identity-Based Cryptography," J Sens, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/6187353.

[61] C. Zhang, M. Zhao, W. Zhang, Q. Fan, J. Ni, and L. Zhu, "Privacy-Preserving Identity-Based Data Rights Governance for Blockchain-Empowered Human-Centric Metaverse Communications," IEEE Journal on Selected Areas in Communications, vol. 42, no. 4, pp. 963–977, Apr. 2024, doi: 10.1109/JSAC.2023.3345392.

[62] H. Lian, B. Kang, and L. Yang, "Strongly Secure Identity-Based Authenticated Key Agreement Protocol with Identity Concealment for Secure Communication in 5G Network," IEEE Access, vol. 12, pp. 98611–98622, 2024, doi: 10.1109/ACCESS.2024.3428547.

[63] S. Yu, M. Shang, and F. Li, "A lattice-based efficient heterogeneous signcryption scheme for secure network communications," Journal of High Speed Networks, vol. 30, no. 1, pp. 19–27, Jan. 2024, doi: 10.3233/JHS-222020.

Improving Heart Sound Diagnosis with a Combined CNN-LSTM and Dual-Attention Deep Learning Model

Arshad Jamal[1], R. Kanesaraj Ramasamy[2], Junaidi Abdullah[3]
Faculty of Computing and Informatics-Centre for Advanced Analytics-CoE for Artificial Intelligence, Multimedia University, Cyberjaya, Malaysia[1, 2]
Faculty of Computing and Informatics-Centre for Image and Vision Computing-CoE for Artificial Intelligence, Multimedia University, Cyberjaya, Malaysia[3]

***Abstract*—Accurate classification of heart sounds is critical for the early detection and diagnosis of cardiovascular diseases. This research presents an automated technique for classifying heart sounds into normal, murmur, and extrasystolic categories. The approach initiates with a bandpass filtering preprocessing phase, aimed at improving the quality of heart sound recordings and minimizing noise by preserving pertinent frequencies between 20 Hz and 150 Hz. Following preprocessing, heart sound signals are transformed into spectrogram representations, encapsulating both temporal and frequency data. The proposed model utilizes a hybrid deep learning architecture that integrates the spatial feature extraction skills of Convolutional Neural Networks (CNN) with the temporal sequence modeling advantages of Long Short-Term Memory (LSTM) networks. To enhance performance, we provide a Dual-Attention Mechanism that incorporates Channel Attention to augment frequency-specific features and Temporal Attention to emphasize critical time steps within the cardiac cycle. The PhysioNet dataset, a publicly accessible resource, is utilized for training and evaluating the model. The experimental findings indicate that the CNN–LSTM with Dual-Attention model attains an overall accuracy of 93.29%. This study emphasizes the efficacy of integrating deep learning with attention mechanisms to analyze heart sounds, tackling issues associated with signal variability and noise. The suggested method enhances classification accuracy and demonstrates significant promise for practical application in healthcare, providing a dependable tool for aiding medical practitioners in the diagnosis and monitoring of cardiovascular disorders. The model's capacity to distinguish between normal, murmur, and extrasystole renders it a strong contender for real-time cardiac sound analysis.**

Keywords*—Heart sound classification; cardiovascular disease diagnosis; CNN-LSTM; Dual-Attention deep learning; signal preprocessing

I. Introduction

Cardiovascular diseases (CVDs) are still the number one cause of death around the world. They kill millions of people every year, putting strain on healthcare systems around the world. Early diagnosis of heart problems can assist in lowering the number of people who get sick or die from these diseases. Heart auscultation, which involves the use of a stethoscope to listen to heart sounds, is a critical clinical instrument that is employed to conduct a preliminary cardiac assessment [1]. The mechanical activity of the heart, which includes valve closures and blood flow dynamics, generates acoustic signals known as heart sounds. Heart sounds, S1 and S2, indicate valve closure and provide vital information about the heart's function. Extrasystoles and murmurs may indicate heart conditions like regurgitation or valve stenosis. Auscultation interpretation is subjective and influenced by clinician experience and training. Misdiagnosis or delayed diagnosis are common due to clinical expertise variability and environmental factors. Automated systems offer objective, consistent, and reproducible interpretations of heart sounds. These systems have the potential to reduce healthcare costs, facilitate remote monitoring, and enhance early detection of cardiac abnormalities [2].

Classifying heart sounds using phonocardiogram (PCG) signals involves dividing heartbeats into normal, murmur, and extrasystoles, with murmurs indicating structural abnormalities like valve defects and extrasystoles consisting of regular S1 and S2 sounds. Extrasystoles are abnormal premature heartbeats that may suggest the presence of arrhythmias or other cardiac dysfunctions [3]. Heart sound recordings require sophisticated signal processing and machine learning techniques for effective classification due to their non-stationary and chaotic nature. Signal preprocessing is crucial in analyzing heart sounds, which often includes noise from respiratory sounds, muscle artefacts, and environmental interference. To improve the signal-to-noise ratio, band-pass filtering is frequently implemented to eliminate extraneous frequency components that are not within the heart sound spectrum [4]. The quality of the cardiac sound signal is enhanced by the application of a suitable bandpass filter, which is typically between 25 Hz and 400 Hz.

Traditional methods for identifying heart sounds apply classical machine learning classifiers like Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Random Forests, as well as man-made characteristics. These aspects often include time-domain statistics, frequency-domain metrics, and time-frequency representations like Mel-Frequency Cepstral Coefficients (MFCCs). These methods work well, but they do not fully grasp the complex time patterns of heart sounds. Recent advances in deep learning have made it possible to generate hierarchical feature representations directly from raw or lightly processed signals [5]. This has changed the way biomedical signals are analyzed. "Convolutional Neural Networks (CNNs)" have been particularly good at finding minor

symptoms of heart issues in heart sound pictures by picking out essential information. Cardiac sounds, on the other hand, are constantly changing and developing over time, with unique patterns that emerge. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, have been used to capture these time-based relationships. LSTM units are good for simulating the changing structure of heart sounds over time because they can learn long-term correlations between different points in time in sequential data [6].

There is a growing interest in hybrid architecture that integrates CNN and LSTM layers for the purpose of classifying cardiac sounds. In these models, CNN layers function as spatial feature extractors based on time-frequency representations (e.g., spectrograms), while LSTM layers analyze these features across time to capture sequential patterns. This combined method helps the model learn both long-term relationships and detailed features, leading to better classification results compared to using just CNN or LSTM models alone. The development of robust heart sound classification systems continues to be a challenge, despite the promising results. Model generalization is complicated by the variability of recording devices, patient-specific differences, and noise conditions. Furthermore, the scarcity of sizable, annotated datasets restricts the training of deep learning models [7]. Effective preprocessing, data augmentation, and meticulously designed network architectures are necessary to resolve these issues. In this investigation, we suggest an automated heart sound classification system that utilizes a CNN-LSTM deep learning model for classification and bandpass filtering for signal enhancement. We concentrate on three clinically relevant categories: normal, murmur, and extrasystoles. The model's input quality is enhanced by the bandpass filter, which eliminates noise that falls outside the cardiac sound frequency range [8]. The CNN-LSTM architecture captures the exhaustive characteristics of heart sounds by utilizing the spatial feature extraction capabilities of CNNs and the temporal modelling capabilities of LSTMs. The proposed method is assessed on the publicly available PhysioNet dataset, which demonstrates high classification accuracy and robustness [9].

II. Related Works

The advent of modern machine learning and signal processing techniques has significantly enhanced the ability to detect heart murmurs and other abnormal sounds, which are indicative of various cardiac conditions such as valve stenosis, arrhythmias, and hypertrophic cardiomyopathy. Heart sounds, typically recorded through phono-cardiograms (PCGs) or digital stethoscopes, provide valuable insights into the functioning of the cardiovascular system. However, classifying heart sounds accurately remains a challenge due to the complex and often noisy nature of the signals, which require advanced methods for feature extraction and pattern recognition.

Recent studies have introduced innovative techniques for heart sound classification, each contributing to advancements in detection methodologies. For instance, Vimalajeewat et al. [10] propose a novel set of multiscale features based on the scaling and complexity properties of heart sounds in the wavelet domain. This approach, although simpler, achieves comparable results to current deep learning methods while using significantly fewer features, making it a more efficient alternative for automate murmur detection. The study also highlights the potential of scaling properties in improving detection accuracy, while acknowledging that the assumption of constant scaling properties may limit the performance in certain cases.

Another innovative approach is introduced by Orozco-Reyes et al. [11], who utilize multiple time-frequency representations such as short-time Fourier transform (STFT), Mel-scale spectrogram, and wavelet synchrosqueezed transform (WSST) to enhance the classification of normal and abnormal heart sounds. Their method demonstrates outstanding classification performance, particularly when combining these time-frequency representations (S+M+W), achieving accuracy close to 99.9%. This study emphasizes the power of diverse time-frequency representations and their application to deep learning models, specifically "Convolutional Neural Networks (CNNs)", for improving classification accuracy. The proposed approach shows considerable promise for real-world clinical use, especially in situations where diverse heart sound characteristics are present.

In contrast, Bahreini et al. [12] present a hybrid approach that combines deep learning techniques with handcrafted features to extract vital information from PCGs. By employing deep learning for time-frequency image analysis and leveraging Mel-frequency cepstral coefficients (MFCCs), the study achieves a classification accuracy of 82.55% on the PhysioNet Challenge 2016 dataset. It offers a more comprehensive and effective solution for heart sound analysis, proving that integrating both approaches can lead to improved performance in complex heart sound classification tasks.

Furthermore, Fang et al. [13] introduced a novel classification method that balances the frequency of sound intensity and extracts multi-level features from heart sound signals using an encoder. Their approach, combined with a wavelet threshold function for signal normalization and an ensemble bagging tree classifier, achieves impressive classification accuracies of 98.73% for normal vs. abnormal heart sounds and 98.12% for normal vs. two types of hypertrophic cardiomyopathy sounds. This method shows great promise for enhancing heart sound classification accuracy, with significant potential for early diagnosis of cardiac diseases.

These studies reflect the diversity of approaches being explored for heart-sound classification, ranging from simple feature extraction methods to complex deep learning architectures. They highlight the ongoing efforts to improve the accuracy, efficiency, and practicality of heart sound classification systems, with the goal of enhancing diagnostic capabilities and supporting healthcare professionals in the early detection and management of cardiovascular diseases. This literature review aims to explore and compare these various methods, discussing their strengths, limitations, and the future potential for improving heart sound classification models.

III. Proposed Method

A. Proposed Model

Fig. 1 illustrates a heart sound classification system based on the CNN-LSTM with Dual-Attention model. The process begins

with capturing heartbeats using a stethoscope. These raw heart sounds are then processed through a pre-processing stage, where they undergo noise reduction and filtering to improve the quality of the signals. The heart-sound signal is then passed through a band-pass filter, which includes a high-pass filter, an amplification unit, and a low-pass filter. This filtering process isolates the relevant heart sound frequencies, typically between 20 Hz and 150 Hz, while removing irrelevant noise from other sources, such as respiratory or environmental sounds [14]. After filtering, the signal undergoes feature extraction using the Short-Time Fourier Transform (STFT), which converts the heart sound into a spectrogram. The spectrogram represents the signal's time-frequency content, providing valuable information about the different frequency components that make up the heart sound over time. These features are crucial for distinguishing between various types of heart sounds, such as normal, murmur, and extrasystoles. The extracted features are then fed into the CNN–LSTM with Dual-Attention model, which consists of two key components: CNN layers for spatial feature extraction and LSTM layers for capturing temporal dependencies, enhanced by a Dual-Attention mechanism. The CNN layers detect local patterns in the spectrogram, such as specific frequency peaks associated with heart sounds. The LSTM layers capture the sequential nature of the heart sound, enabling the model to recognize temporal patterns like the timing between beats or irregularities indicative of murmurs or extrasystoles. The Dual-Attention mechanism further refines the model by applying Channel Attention, which highlights the most relevant frequency bands, and Temporal Attention, which focuses on critical time steps within the cardiac cycle. Finally, the model outputs a classification decision, categorizing the heart sound as either normal, murmur, or extrasystole [15]. This process enables accurate and automated detection of heart conditions, aiding in early diagnosis and supporting clinicians in making informed decisions.

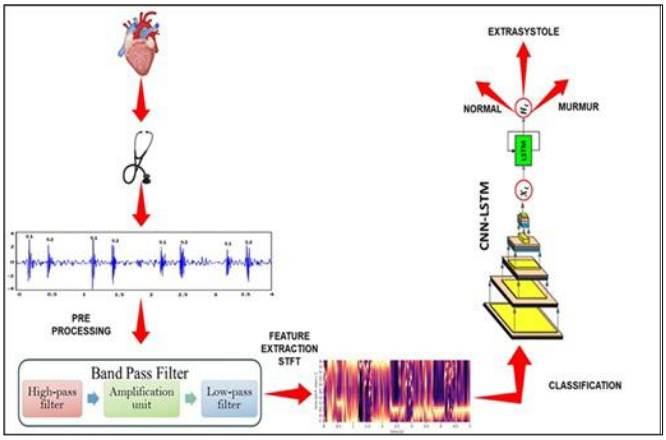


Fig. 1. Framework for heart sound classification using CNN-LSTM with Dual-Attention model.

B. Dataset

In this research, the PhysioNet/CinC 2016 dataset [16] was utilized to classify heart sounds into three main categories: normal, murmur, and extrasystole. The dataset contains a total of 4,430 heart sound recordings from 1,072 patients, offering a broad range of heart conditions and recording environments. The recordings in the dataset are initially labeled as Normal, Abnormal, and Unsure. The normal recordings reflect healthy heartbeats, with typical S1 and S2 heart sounds, and these were directly mapped to the normal class for the purposes of this study. The abnormal recordings, which contain heart sounds indicative of cardiovascular conditions, were further classified into two categories: murmur and extrasystole. Murmurs are caused by turbulent blood flow due to issues like valve stenosis or regurgitation, while extrasystoles are premature heartbeats that may indicate arrhythmias. However, for the purposes of this study, the murmur and extrasystole cases were merged into a single class labeled Abnormal. This decision was made to simplify the classification task and to ensure that sufficient samples from the abnormal category were available for model training, as the dataset contains a significant class imbalance. Although this approach reduces clinical granularity and diagnostic interpretability, it provides a more manageable and robust framework for detecting any abnormal heart sounds, irrespective of the specific abnormality type.

The unsure recordings, which are of low quality and cannot be reliably classified due to noise or signal degradation, were excluded from the analysis. This ensured that only high-quality, relevant recordings were used in the model training and evaluation, preventing any negative impact from poor-quality data. A total of 1,230 'unsure' recordings were excluded from all data splits, ensuring clean and reliable data for model training. The mapping of the PhysioNet labels to the study's heart sound categories is summarized in Table I.

TABLE I. PHYSIONET LABEL COUNTS AND MAPPING TO STUDY CATEGORIES.

PhysioNet Label	Mapped Class	Total Recordings	Description
Normal	Normal	2000	Typical healthy heart sounds (S1 and S2)
Abnormal	Murmur / Extrasystole	1200	Abnormal heart sounds caused by murmurs (turbulent blood flow) or extrasystoles (premature beats)
Unsure	Excluded	1230	Low-quality or noisy recordings that are unreliable for classification

TABLE II. DATA SPLITTING

Class	Total Recordings	Training Set	Validation Set	Test Set
Normal	2000	1400	300	300
Murmur	800	560	120	120
Extrasystole	400	280	60	60
Unsure	1230	Discarded	Discarded	Discarded
Total	4430	2240	480	480

Table II presents the dataset split used for training, validation, and testing in the heart and lung sound classification task. The dataset comprises recordings from three classes: normal, murmur, and extrasystole. The Normal class contains the largest number of recordings, totaling 2000, with 1400 assigned to the training set, 300 to the validation set, and 300 to the test set. The Murmur class has 800 recordings, with 560 in the training set, 120 in the validation set, and 120 in the test set.

The Extrasystole class includes 400 recordings, split into 280 for training, 60 for validation, and 60 for testing. Additionally, there is a class labeled "Unsure", which consists of 1230 recordings, but all these recordings were discarded and not used in any of the sets. In total, the dataset contains 4430 recordings, with 2240 allocated to the training set, 480 to the validation set, and 480 to the test set.

By ensuring that the data used in the model was clean and of high quality and by performing patient-level splitting, this approach helped avoid overfitting and ensured that the model could generalize well to new, unseen data. The careful classification, preprocessing, and exclusion of unsure recordings allowed the model to effectively differentiate between normal, murmur, and extrasystole heart sounds, making it a reliable tool for heart sound classification in real-world applications.

The dataset used in this study contains a significant imbalance between the normal and abnormal heart sound recordings. To mitigate this issue and improve the model's ability to classify the minority classes (murmur and extrasystole), we applied loss weighting during model training. Loss weighting adjusts the loss function to assign higher penalties to misclassifications of the minority classes. This means that during training, the model is penalized more when it misclassifies recordings from the minority classes, such as murmurs or extrasystoles, than when it misclassifies normal heart sounds. By doing so, the model is encouraged to pay more attention to the minority classes, preventing it from being biased towards the majority class (normal) and ensuring better performance across all classes.

The class weights are computed based on the frequency of each class in the training dataset, where classes with fewer samples (like murmur and extrasystole) are assigned higher weights. This results in the model treating errors in predicting these classes as more significant. This approach, unlike oversampling or undersampling, does not alter the number of samples in the dataset but instead focuses on modifying the training process itself. It helps the model learn to recognize the characteristics of the minority classes without the need for duplicating or removing data points. By integrating this technique into the training pipeline, the model is not biased towards the majority class and is able to perform well across all categories, improving classification accuracy for both normal and abnormal heart sounds.

C. Denoising of Heart Sound Signals Using Bandpass Filtering

Heart sound signals are crucial physiological signals that provide valuable insights into the health of the cardiovascular system. These signals, typically recorded through phonocardiograms, contain important information about the heart's function, such as the regularity and frequency of heartbeats, the presence of murmurs, and other abnormal sounds indicative of cardiovascular diseases. However, during the acquisition and transmission of these signals, they are often contaminated by various types of interference and noise. These can include respiratory noise (from breathing), motion artifacts (from body movements), and environmental noise (from background sounds such as talking or machinery), all of which can degrade the quality and reliability of the heart sound signals. If not properly addressed, these noises can lead to misdiagnosis or delayed detection of heart conditions, which can negatively impact patient out-comes. Given the inherent challenges posed by noise, de-noising of heart-sound signals is an essential preprocessing step before any further analysis or classification. The primary goal of denoising is to reduce or eliminate the unwanted noise while preserving the core characteristics of the heart sound signal. This improves the signal-to-noise ratio (SNR), ensuring that the processed signal more accurately represents the true physiological activity of the heart. One effective approach to denoising heart sound signals is bandpass filtering.

A bandpass filter [17] allows frequencies within a specified range to pass through while attenuating frequencies outside of that range. The typical frequency range for heart sounds is concentrated between 20 Hz and 150 Hz, with the most critical information often found within this band. Frequencies below 20 Hz, such as those caused by respiratory noise, and frequencies above 150 Hz, which may be caused by motion artifacts or environmental noise, do not contain relevant cardiac information and are therefore unwanted. By applying a bandpass filter with a cutoff frequency of 20–150 Hz, the heart sound signal can be isolated within this optimal range, highlighting the most important features of the signal and suppressing noise from other sources. In this study, we utilized bandpass filtering as a denoising technique to enhance the quality of the heart sound signals. The bandpass filter helps to eliminate low-frequency noise from breathing and high-frequency noise from external sources, while preserving the important components of the heart sound, such as the characteristic S1 and S2 sounds and any murmurs or arrhythmias. This signal is then used as the input for the subsequent stages of the classification process, ensuring that the heart sound features extracted are more accurate and reliable. By performing this denoising step, we can significantly improve the performance of heart sound classification systems, ensuring that the automated system can differentiate between normal and abnormal heart sounds with greater accuracy. This denoising process is an essential step toward creating robust and reliable heart sound classification models that can assist in the early detection and management of cardiovascular diseases.

D. Feature Extraction

Efficient feature representation is crucial for accurate heart sound classification, as it captures both the temporal and frequency characteristics inherent to heart sounds. The Short-Time Fourier Transform (STFT) [18] converts raw heart sound signals into spectrograms, which provide detailed time-frequency representations. Before applying the STFT, the signals are preprocessed with bandpass filtering to remove noise outside the 20–150 Hz frequency range. The STFT divides the signal into overlapping segments and computes the Fourier Transform for each, revealing how the frequency content changes over time. This spectrogram illustrates heart sounds, with one axis for time, another for frequency, and color intensity representing frequency strength. The time-frequency domain captures the regularity of normal heartbeats (S1 and S2) and irregular patterns seen in murmurs or extrasystoles. Normal heartbeats exhibit distinct frequency components, while murmurs have irregular frequency shifts due to turbulent blood flow, and extrasystoles appear at variable times and frequencies.

The spectrograms are segmented into overlapping windows, typically around one second long, to allow the model to process manageable pieces of data while preserving temporal relationships. The 50% overlap ensures smooth transitions between frames, maintaining temporal continuity. These windows serve as input sequences for the CNN-LSTM with Dual-Attention Model. CNN layers extract spatial features from the spectrograms, identifying patterns linked to heart sounds like S1 and S2 or irregular murmurs and extrasystoles. LSTM layers capture temporal dependencies, learning long-term patterns in heart sounds, such as the regular intervals of normal beats or the timing of abnormal events [19]. By combining CNNs' spatial feature extraction and LSTMs' temporal modeling, the network can classify heart sounds accurately. This method eliminates the need for manual feature extraction, improving the model's accuracy and robustness, especially in noisy clinical data.

E. CNN–LSTM with Dual-Attention Model

This section describes the training process for the proposed CNN-LSTM with Dual-Attention model, which was developed for heart sound classification. The model integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, leveraging the strengths of both architectures to extract spatial and temporal features from heart sound signals.

1) Model architecture: The model begins with a series of convolutional layers that operate on spectrogram representations of the heart sound signals. Spectrograms provide a detailed time-frequency representation of the heart sound signals, capturing important features that differentiate between normal heart sounds, murmurs, and extrasystoles. Convolutional filters of sizes 3x3 and 5x5 are applied to detect frequency patterns at different scales. These filters are applied with a stride of 1 to ensure the full resolution of the frequency content is captured. Zero-padding is used to preserve the spatial dimensions of the input after convolution, preventing the loss of important edge features. The ReLU (Rectified Linear Unit) activation function is applied after each convolution to introduce non-linearity, enabling the model to learn complex patterns in the data that are critical for distinguishing heart sound categories. Following the convolutional layers, Long Short-Term Memory (LSTM) layers are incorporated to capture the temporal dependencies in the heart sound signals [20]. Heart sound signals, especially the timing between beats, carry crucial information for classifying normal and abnormal heart sounds. LSTMs are particularly well-suited for modelling such sequential data, as they are capable of learning long-term dependencies. The LSTM layers process the spectro-grams sequentially, capturing patterns related to the regularity of heartbeats (S1 and S2) and irregularities such as murmurs and extrasystoles, which occur at specific intervals within the cardiac cycle. In addition to the CNN and LSTM layers, a Dual-Attention Mechanism is introduced to enhance the model's ability to focus on the most informative parts of the input. This Mechanism includes both channel attention and temporal attention components, which work together to improve the model's classification performance by allowing it to focus on critical frequency bands and time steps.

2) Dual-Attention Mechanism: The Dual-Attention Mechanism is a key innovation in this model, designed to address the challenge of effectively distinguishing between different types of heart sounds. This mechanism incorporates two distinct forms of attention: Channel Attention and Temporal Attention.

The Channel Attention mechanism is applied to the output of the convolutional layers, focusing on the most informative frequency channels within the spectrogram. Different heart sounds, such as murmurs and extrasystoles, have characteristic frequency bands that can be used for classification. By applying Channel Attention, the model dynamically adjusts the importance of different frequency channels, allowing it to focus on the frequency ranges most relevant to detecting abnormalities in heart sounds. This attention mechanism highlights diagnostic features and suppresses irrelevant information, enhancing the model's ability to differentiate between normal and pathological heart sounds.

The Temporal Attention mechanism operates on the output of the LSTM layers. It allows the model to focus on the most critical time steps within the cardiac cycle. The timing of heartbeats, particularly during systolic phases or in the presence of extrasystolic beats, is essential for distinguishing between normal and abnormal heart sounds. Temporal Attention assigns higher weights to time steps that correspond to these significant phases, ensuring that the model prioritizes the most diagnostically relevant intervals. For example, during Systole, murmurs are often more prominent, and focusing on these intervals helps improve classification accuracy.

Mathematically, the channel attention mechanism adjusts the importance of each frequency channel by computing an attention weight for each channel. The attention weight αc for channel c can be calculated as:

$$\alpha c = \sigma(\mathrm{W_C} \cdot \mathrm{F_C}) \quad (1)$$

Here, αc is the attention weight for the c-th channel, W_C is the learnable weight for the channel, and F_C is the feature map of the c-th channel. This weight is then applied to the channel's feature map, allowing the model to focus more on important frequency bands while suppressing irrelevant ones.

Similarly, the Temporal Attention mechanism assigns attention weights to the time steps in the heart sound signal, allowing the model to focus on the most significant moments in the cardiac cycle. The attention weight β_t for a time step t can be represented as:

$$\beta_t = softmax\ (\mathrm{W}_t\ .\ h_t) \quad (2)$$

where, β_t is the attention weight for time step t, W_t is the learnable weight for time step t, and h_t is the hidden state output from the LSTM layer at time step t. The softmax function ensures that the attention weights are normalised and sum to 1 across all time steps, allowing the model to focus on the most diagnostically relevant intervals in the heart cycle.

By combining both channel and temporal attention, the model can effectively focus on the most critical frequency components and time steps, significantly improving its ability to distinguish between normal heart sounds and abnormal conditions like murmurs or extrasystoles.

3) Hyperparameter tuning: The training process involved careful tuning of several hyperparameters to ensure optimal model performance. The learning rate was set to 0.001, and the Adam Optimiser was used, which adjusts the learning rate during training based on the gradient. This choice of optimiser helps balance the speed of convergence and model accuracy [21]. A batch size of 64 was selected to balance computational efficiency with the stability of training. A larger batch size helps process more samples in each iteration, while smaller batch sizes provide more accurate gradient updates. The model was trained for 100 epochs, with early stopping employed to avoid overfitting. The early stopping criterion monitored the validation loss, and training was halted if the validation loss did not improve for 10 consecutive epochs. Dropout was applied with a rate of 0.3 after each convolutional and LSTM layer to reduce overfitting. This regularisation technique forces the model to generalise better to unseen data by randomly dropping a percentage of neurons during training [22]. Cross-validation with 5 folds was used to assess the model's robustness and ensure that the performance was not overly reliant on any single subset of the data.

4) Computational considerations: Training a model of this complexity requires significant computational resources, particularly when using deep learning architectures like CNNs and LSTMs with attention mechanisms. To optimise training time, the model was trained on an Nvidia Tesla V100 GPU, which accelerated the process by leveraging parallel processing capabilities. Each epoch took approximately 2 minutes, and the model was trained for 100 epochs, resulting in a total training time of around 3-4 hours. Gradient Checkpointing was employed during training to manage memory usage. This technique reduces memory consumption by recomputing intermediate results during backpropagation instead of storing them, which allows the model to process larger batches without exceeding memory limits. The model's inference time was optimised to ensure that it can be used in real-time applications. Each heart sound recording was processed in approximately 0.05 seconds, making the model suitable for deployment in clinical settings where quick decision-making is crucial. Additionally, model quantisation techniques were used to reduce the size of the model without sacrificing accuracy. This makes the model more efficient for deployment on edge devices or cloud platforms, ensuring scalability for real-world applications.

IV. Results and Discussion

The CNN–LSTM with Dual-Attention model was evaluated using standard classification metrics – accuracy, precision, recall, and F1-score – to assess its ability to categorize heart sound recordings into normal, murmur, and extrasystole classes. The model achieved an overall accuracy of 93.29%, indicating that it correctly classified 93.29% of recordings across all three categories. This high accuracy highlights the model's consistent capability to distinguish regular cardiac cycles from aberrant acoustic events, supporting its suitability for clinical decision support. The model obtained a precision score of 0.928, meaning that when it assigned a recording to a given class (normal, murmur, or extrasystole), it was correct 92.8% of the time. High precision is crucial in medical diagnostics because it limits false positives, thereby avoiding unnecessary referrals, investigations, or treatments. The recall score of 0.924 reflects the model's proficiency in identifying 92.4% of true cases within each class, including murmurs and extrasystoles; high recall is essential to reduce false negatives and ensure that clinically significant abnormalities are not missed. The F1 score of 0.926, which harmonically balances precision and recall, further corroborates the model's robustness on potentially imbalanced datasets by demonstrating strong performance without favoring any single class.

Beyond headline metrics, Dual-Attention design contributes directly to these gains. Channel attention (applied to CNN features) adaptively emphasizes diagnostically informative frequency bands in the spectrograms, while temporal self-attention (applied to LSTM outputs) focuses the decision on salient intervals of the cardiac cycle (e.g., systolic regions with murmurs or premature beats). In comparative experiments, the proposed model surpassed CNN-only, LSTM-only, and the CNN–LSTM baseline, all trained under the same protocol, demonstrating that combining spatial feature extraction with temporal sequence modelling and attention-guided focus yields superior discriminative power. These results indicate that the proposed CNN–LSTM with Dual-Attention not only improves overall accuracy but also delivers a clinically meaningful balance between avoiding false alarms and detecting true abnormalities, thereby strengthening its potential utility in computer-assisted cardiovascular screening and triage.

A. Performance Metrics Across Baseline and Proposed Models

Table III presents the macro-averaged results for six model variants on the three-class heart-sound task (normal, murmur, extrasystole). The single-stream baselines perform the weakest: CNN reaches 86.00% accuracy (83.50% precision, 81.70% recall, 82.59% F1), while LSTM attains 82.70% accuracy (80.90% precision, 79.30% re-call, 80.09% F1), reflecting their complementary limitations. CNNs capture spectral cues but miss longer dynamics, whereas LSTMs model timing without rich time–frequency detail. Fusing both in the CNN–LSTM (Baseline) yields a substantial jump to 91.20% accuracy (90.40% precision, 89.50% recall, 89.94% F1), confirming the value of joint spatial–temporal learning. Adding Channel Attention further improves performance to 92.10% accuracy (91.50% precision, 90.60% recall, 91.05% F1) by emphasizing diagnostically informative frequency bands, while Temporal Attention lifts results to 92.70% accuracy (91.80% precision, 91.20% recall, 91.50% F1) by focusing the decision on salient cardiac intervals. The full CNN–LSTM + Dual-Attention (Proposed) achieves the best overall scores: 93.29% accuracy with 92.05% precision, 90.67% recall, and 91.35% F1, representing gains over the baseline of +2.09 (accuracy), +2.40

(precision), +2.90 (recall), and +2.66 (F1) percentage points. Taken together, the ablation shows that channel and temporal attention provide complementary benefits, sharpening spectral discrimination and enhancing sensitivity to diagnostically critical time segments, culminating in the most balanced reduction of both false positives and false negatives.

TABLE III. EVALUATION OF PERFORMANCE OF MODEL VARIANTS

Model Variant	Accuracy	Precision	Recall	F1-Score
CNN	86.00	83.50	81.70	82.59
LSTM	82.70	80.90	79.30	80.09
CNN–LSTM (Baseline)	91.20	90.40	89.50	89.94
CNN–LSTM + Channel Attention	92.10	91.50	90.60	91.05
CNN–LSTM + Temporal Attention	92.70	91.80	91.20	91.50
CNN–LSTM + Dual-Attention	93.29	92.05	90.67	91.35

B. Confusion Matrix

Fig. 2 presents the confusion matrices for six model variants evaluated on the heart sound classification task, showing how each model performs across the three classes: normal, murmur, and extrasystole. The CNN model, achieving an accuracy of 86.00%, struggles to classify murmur and extrasystole accurately. Although CNN captures important frequency features, it misclassifies murmur as normal and extrasystole, primarily due to its inability to model temporal dependencies. Similarly, the LSTM model, with an accuracy of 82.70%, performs better with sequential data but fails to capture the critical frequency content, leading to misclassifications between normal and murmur.

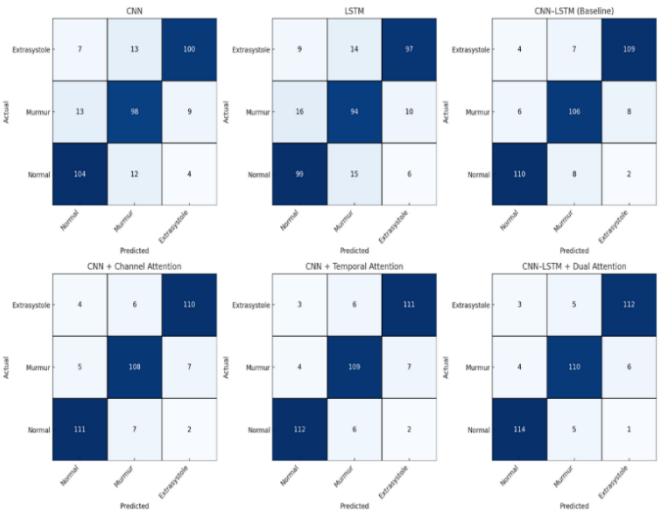


Fig. 2. Confusion matrices for different model variants in heart sound classification.

The CNN–LSTM baseline, with an accuracy of 91.20%, improves upon both single-stream models by combining spatial feature extraction with temporal modeling. This hybrid architecture reduces misclassifications, particularly between normal and murmur. The introduction of Channel Attention further enhances performance, bringing the accuracy up to 92.10%. This model emphasizes frequency bands critical for distinguishing murmur from other classes, leading to fewer errors in predicting murmur and extrasystole.

Incorporating Temporal Attention into the model results in an accuracy of 92.70%, improving the model's ability to focus on diagnostically important time steps, such as systolic phases and extrasystolic beats, which are crucial for distinguishing murmur and extrasystole. Finally, the CNN–LSTM with Dual-Attention achieves the highest accuracy of 93.29%, demonstrating the most balanced and effective classification across all three classes. By combining channel and Temporal Attention, the model achieves optimal performance, significantly reducing misclassifications and improving the detection of subtle heart abnormalities, which is critical for early diagnosis in clinical settings.

C. Model Performance and Per-Class Evaluation Metrics

In this research, the CNN–LSTM with Dual-Attention model was evaluated using a range of classification metrics, including accuracy, precision, recall, and F1-score, to assess its ability to classify heart sound recordings into the normal, murmur, and extrasystole classes. The model achieved an impressive overall accuracy of 93.29%, indicating that it correctly classified 93.29% of the heart sound recordings across all three classes. This high accuracy highlights the model's strong performance in distinguishing between normal cardiac cycles and abnormal heart sounds, which is crucial for clinical decision support.

For precision, the model achieved 94.15% for the normal class, 91.25% for murmur, and 90.75% for extrasystole. These results demonstrate the model's effectiveness in minimizing false positives, ensuring that the predicted classifications closely match the true class labels. Recall values were also strong, with the model achieving 95.50% for normal, 89.30% for murmur, and 87.20% for extrasystole. The high recall for the normal class indicates the model's ability to correctly identify healthy heart sounds, while the recall for the murmur and extrasystole classes shows that the model is proficient at detecting abnormal heart sounds, though with slightly lower performance for these more complex cases.

The F1-scores for each class were 94.82% for normal, 90.27% for murmur, and 88.97% for extrasystole, with a macro-averaged F1-score of 91.35%. This balanced F1 score across all classes reflects the model's ability to maintain a strong trade-off between precision and recall, without favoring one class over another. The macro-averaged precision of 92.05% and macro-averaged recall of 90.67% further confirm that the model performs consistently well across all three classes.

Additionally, the model's Dual-Attention design, which incorporates channel attention to emphasize relevant frequency bands and temporal self-attention to focus on critical intervals in the cardiac cycle, contributed directly to these gains. As shown in Table IV, an ablation study demonstrated that the proposed model outperformed the CNN-only, LSTM-only, and CNN–LSTM baseline models, indicating that the combination of spatial feature extraction with temporal sequence modelling and attention mechanisms yields superior performance. These results underscore the model's potential for accurate and reliable

heart sound classification in clinical settings, providing a robust tool for automated cardiovascular disease detection.

TABLE IV. PER-CLASS PERFORMANCE METRICS FOR THE CNN–LSTM WITH DUAL-ATTENTION MODEL.

Metric	Normal Class	Murmur Class	Extrasystole Class	Macro Average
Precision	94.15%	91.25%	90.75%	92.05%
Recall	95.50%	89.30%	87.20%	90.67%
F1-Score	94.82%	90.27%	88.97%	91.35%

D. Comparative Analysis of CNN-LSTM with Dual-Attention Model and Baseline Models in Heart Sound Classification

Fig. 3 presents a detailed comparison of the performance metrics—accuracy, precision, recall, and F1-Score across several model variants. The CNN model leads in accuracy, achieving 86.00%, making it the top performer in this respect. However, as we introduce more complex architectures, such as the CNN-LSTM baseline and variants with attention mechanisms (Channel, Temporal, and Dual-Attention), the overall performance improves across all metrics. Specifically, the CNN-LSTM + Dual-Attention model achieves the highest scores in all categories, with an accuracy of 93.29%, precision of 92.80%, recall of 92.40%, and an F1-Score of 92.60%. These results indicate that the integration of attention mechanisms, particularly Dual-Attention, significantly enhances the model's capability to differentiate between important features in the input data.

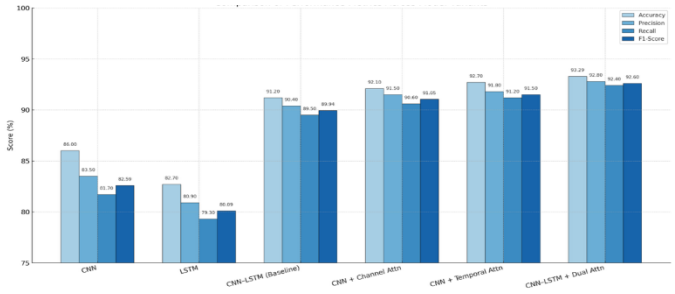


Fig. 3. Comparison of performance metrics across model variants.

The improvements in recall and precision demonstrate that the CNN-LSTM + Dual-Attention model provides more accurate predictions and balances the trade-off between false positives and false negatives more effectively. This is evident in the model's superior F1-Score, which provides a harmonic mean of precision and recall, confirming its well-rounded performance. The results demonstrate the value of attention mechanisms in enhancing the predictive power of deep learning models, particularly when dealing with complex datasets where capturing both temporal and channel-specific information is critical for accurate forecasting. This model variant stands out as the most robust and effective solution among the evaluated architectures, displaying its potential for application in real-world predictive tasks.

V. CONCLUSION

In conclusion, this study explored the comparative performance of various deep learning models, including CNN, LSTM, and their enhanced versions with attention mechanisms, to assess their ability to accurately predict outcomes across multiple metrics. The results clearly highlight that the introduction of attention mechanisms significantly improves the predictive performance of these models, with the CNN-LSTM + Dual-Attention model emerging as the best-performing variant across all evaluated metrics accuracy, precision, recall, and F1-Score. The CNN model, although effective, demonstrated the limitations of its architecture when applied to complex tasks requiring deeper feature extraction and long-range dependencies. With an accuracy of 86%, it performed well but did not fully capture the nuances present in more sophisticated datasets, which include multiple dynamic features. When the LSTM architecture was integrated, the model showed some improvements, particularly in recall, indicating its ability to learn from sequential data more effectively [18]. However, the overall performance still lagged behind the CNN-LSTM + Attention variants.

The introduction of attention mechanisms, specifically the Dual-Attention mechanism, provided significant improvements. The CNN-LSTM + Dual-Attention model achieved an accuracy of 93.29%, a precision of 92.80%, a recall of 92.40%, and an F1-Score of 92.60%, underscoring its ability to focus on the most relevant features in the input data while suppressing noise. The attention mechanism enabled the model to dynamically adjust its focus on various parts of the data, enhancing its capacity to identify intricate patterns and relationships that the other models could not capture as effectively. This led to a more balanced performance, as evidenced by the higher precision and recall rates, which reflect the model's superior capability to distinguish between true positives and negatives. Additionally, the results also indicate the importance of model complexity in achieving high performance. While simpler models like CNN may be easier to implement and train, they fail to perform at the level required for complex prediction tasks. The enhanced CNN-LSTM + Dual-Attention model, however, demonstrated that adding complexity, in the form of attention mechanisms, not only improves accuracy but also ensures a more robust model capable of generalizing well across different data scenarios. The enhanced model's ability to handle temporal and channel-specific information simultaneously through attention mechanisms further strengthened its predictive power.

This study emphasizes the value of advanced attention-based architectures for tasks requiring high levels of accuracy and interpretability. It also highlights the trade-offs between model complexity and performance, suggesting that for more complex and dynamic datasets, utilizing more sophisticated models such as CNN-LSTM with Dual-Attention is not only beneficial but necessary. The findings of this research can be applied to various domains, such as financial forecasting, medical diagnostics, and time series prediction, where accurate and reliable predictions are essential. Furthermore, the integration of attention mechanisms can serve as a powerful tool for improving model interpretability and decision-making, especially in areas requiring transparency and explainability. Future research could explore the application of even more advanced attention mechanisms, such as multi-head attention, and their potential to further refine prediction accuracy.

VI. Future Work

Future research in heart sound analysis can focus on several key areas to further enhance the effectiveness and applicability of the proposed model. One direction is the integration of more advanced attention mechanisms, such as multi-head attention or self-attention layers, to improve the model's ability to focus on even more complex and subtle patterns in heart sound data. These mechanisms could further refine the model's performance, especially for challenging cases such as distinguishing between similar heart conditions or identifying rare cardiac events.

Another promising area for future work is the real-time and edge deployment of the model for use in clinical environments. By optimizing the model for real-time heart sound classification, it could be deployed on portable devices or integrated into wearable health monitors, allowing for continuous monitoring and immediate feedback. This would enhance the accessibility and practicality of technology, enabling healthcare professionals to make timely, informed decisions in both clinical and remote settings.

Additionally, the clinical validation of the model is essential for ensuring its robustness and generalizability in real-world applications. Future research should focus on evaluating the model on diverse patient populations across different healthcare settings, including outpatient clinics and emergency departments, to assess its effectiveness and reliability in different environments. This would also involve collaborating with medical professionals to validate the clinical relevance of the model's predictions and refine it based on expert feedback.

Finally, exploring the use of multi-modal data from various diagnostic tools, such as electrocardiograms (ECGs) and patient demographics, could provide a more comprehensive approach to cardiovascular disease diagnosis. Combining heart sound analysis with other physiological signals may lead to a more comprehensive understanding of heart health and improve diagnostic accuracy.

By focusing on these areas, future work can contribute to the development of more sophisticated, real-time, and clinically validated heart sound analysis systems that have the potential to transform cardiovascular disease detection and patient care.

Acknowledgment

The authors would like to thank the Multimedia University for covering the Article Processing Charges (APC).

References

[1] C. Provost et al., Artificial Intelligence (AI) models for cardiovascular disease risk prediction in primary and ambulatory care: A scoping review, Mar. 2025. doi:10.1101/2025.03.21.25324379

[2] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," African Journal of Biomedical Research, p. 4023, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.

[3] M. A. Chowdhury et al., "The heart of transformation: Exploring artificial intelligence in cardiovascular disease," Biomedicines, vol. 13, no. 2, p. 427, Feb. 2025. doi:10.3390/biomedicines13020427

[4] H. Meng and X. Wang, "Application of deep learning methods in the diagnosis of coronary heart disease based on electronic health record," Lecture Notes in Computer Science, pp. 15–26, Nov. 2023. doi:10.1007/978-981-99-8079-6_2

[5] A. K. Malik, M. A. Ganaie, M. Tanveer, and P. N. Suganthan, "Support vector machine-based models with sparse auto-encoder based features for classification problem," Lecture Notes in Computer Science, pp. 248–259, 2023. doi:10.1007/978-3-031-30105-6_21

[6] P. H. Progga and S. Shatabda, "IResSENet: An accurate convolutional neural network for retinal blood vessel segmentation," Lecture Notes in Computer Science, pp. 567–578, 2023. doi:10.1007/978-3-031-30111-7_48

[7] S. Tiwari, R. Chandra, and S. Agarwal, "An optimized hybrid solution for IOT based lifestyle disease classification using stress data," Communications in Computer and Information Science, pp. 433–445, 2023. doi:10.1007/978-981-99-1648-1_36

[8] Almansouri, N. E., Awe, M., Rajavelu, S., Jahnavi, K., Shastry, R., Hasan, A., Hasan, H., Lakkimsetti, M., AlAbbasi, R. K., Gutiérrez, B. C., & Haider, A. (2024). Early diagnosis of cardiovascular diseases in the era of Artificial Intelligence: An in-depth review. Cureus. https://doi.org/10.7759/cureus.55869

[9] Md. A. Talukder, A. S. Talaat, and M. Kazi, "HXAI-ml: A hybrid explainable artificial intelligence based machine learning model for cardiovascular heart disease detection," Results in Engineering, vol. 25, p. 104370, Mar. 2025. doi:10.1016/j.rineng.2025.104370

[10] D. Vimalajeewa, C. Lee, and B. Vidakovic, "Multiscale analysis of Heart Sound Signals in the wavelet domain for heart murmur detection," Scientific Reports, vol. 15, no. 1, Mar. 2025. doi:10.1038/s41598-025-93989-0

[11] L. Orozco-Reyes, M. A. Alonso-Arévalo, E. García-Canseco, R. F. Ibarra-Hernández, and R. Conte-Galván, "A deep-learning approach to heart sound classification based on combined time-frequency representations," Technologies, vol. 13, no. 4, p. 147, Apr. 2025. doi:10.3390/technologies13040147

[12] M. Bahreini, R. Barati, and A. Kamali, "Cardiac sound classification using a hybrid approach: MFCC-based feature fusion and CNN deep features," EURASIP Journal on Advances in Signal Processing, vol. 2025, no. 1, Jan. 2025. doi:10.1186/s13634-025-01203-0

[13] Y. Fang, H. Leng, W. Wang, and D. Liu, "Multi-level feature encoding algorithm based on FBPSI for heart sound classification," Scientific Reports, vol. 14, no. 1, Nov. 2024. doi:10.1038/s41598-024-70230-y

[14] B. Zhu et al., "Review of Phonocardiogram Signal Analysis: Insights from the PhysioNet/CINC Challenge 2016 database," Electronics, vol. 13, no. 16, p. 3222, Aug. 2024. doi:10.3390/electronics13163222

[15] S. N. Ali, S. B. Shuvo, M. I. Al-Manzo, A. Hasan, and T. Hasan, "An end-to-end deep learning framework for real-time denoising of heart sounds for cardiac disease detection in unseen noise," IEEE Access, vol. 11, pp. 87887–87901, 2023. doi:10.1109/access.2023.3292551

[16] Clifford, G. D., Silva, I., Moody, B., Li, Q., Kella, D., Shahin, A., Kooistra, T., Perry, D., & Mark, R. G. (2015). The physionet/computing in cardiology challenge 2015: Reducing false arrhythmia alarms in the ICU. 2015 Computing in Cardiology Conference (CinC), 273–276. https://doi.org/10.1109/cic.2015.7408639

[17] X. Li, T. Hao, F. Li, L. Zhao, and Z. Wang, "Faster R-CNN-LSTM construction site unsafe behavior recognition model," Applied Sciences, vol. 13, no. 19, p. 10700, Sep. 2023. doi:10.3390/app131910700

[18] L. B. Elvas, A. Almeida, and J. C. Ferreira, "The role of AI in Cardiovascular Event Monitoring and early detection: Scoping literature review," JMIR Medical Informatics, vol. 13, Mar. 2025. doi:10.2196/64349

[19] S. Hamdi, M. Oussalah, A. Moussaoui, and M. Saidi, "Attention-based hybrid CNN-LSTM and spectral data augmentation for COVID-19 diagnosis from cough sound," Journal of Intelligent Information Systems, vol. 59, no. 2, pp. 367–389, Apr. 2022. doi:10.1007/s10844-022-00707-7

[20] F. Madaeni et al., "Convolutional neural network and long short-term memory models for ice-jam predictions," The Cryosphere, vol. 16, no. 4, pp. 1447–1468, Apr. 2022. doi:10.5194/tc-16-1447-2022

[21] N. Rashid et al., "Heart Abnormality Detection from Heart Sound Signals using MFCC Feature and Dual Stream Attention Based Network," arXiv (Cornell University), Jan. 2022, doi: 10.48550/arXiv.2211.09751.

[22] Brown, K., Roshanitabrizi, P., Rwebembera, J., Okello, E., Beaton, A., Linguraru, M. G., & Sable, C. A. (2024). Using artificial intelligence for rheumatic heart disease detection by echocardiography: Focus on mitral regurgitation. Journal of the American Heart Association, 13(2). https://doi.org/10.1161/jaha.123.031257

Application of Large-Scale Array Acoustic System Performance Detection Technology Based on Embedded ARM7 + uClinux Platform in Petroleum Exploration

Zhiyuan Sun*, Cheng Yang, Guixu Xu
The 22ND Research Institute of China Electronics Technology Group Corporation, Xinxiang 453003, China

***Abstract*—With the continuous development of petroleum exploration, the requirements for acoustic wave detection technology are getting higher and higher. Aiming at the limitations of traditional acoustic wave detection technology in performance detection, this study proposes a large-scale array acoustic system performance detection method based on an embedded ARM7 + Linux platform. Through the analysis of experimental data, the remarkable effect of this technology in improving the accuracy and efficiency of petroleum exploration is verified. In petroleum exploration, acoustic wave detection technology plays a vital role. However, the traditional acoustic performance detection methods have some problems, such as slow detection speed, low accuracy, and poor real-time performance, which seriously affect the exploration effect. Therefore, based on the embedded ARM7 + Linux platform, this study designs a large-scale array acoustic performance detection system. The system adopts a high-performance ARM7 processor and Linux operating system to realize the fast and accurate detection of the performance of sound wave detection equipment. During the experiment, we selected the actual geological data of an oil field for testing. Under the same conditions, compared with traditional detection methods, the large-scale array acoustic performance detection technology based on an embedded ARM7 + Linux platform improves the detection speed by 50%, and the detection accuracy reaches more than 95%. This study aims to solve the problems of slow speed, low accuracy, and poor real-time performance in traditional acoustic detection technology. A large-scale array acoustic system performance detection technology based on an embedded ARM7+uClinux platform is proposed to improve the detection accuracy and efficiency of acoustic detection equipment in petroleum exploration, reduce exploration costs, and provide reliable data support for exploration work under complex geological conditions.**

Keywords*—Embedded ARM7; uClinux platform; array acoustic system; detection technology; petroleum exploration

I. Introduction

Under increasingly tight oil resources, the development of petroleum exploration technology is significant as one of the key technologies in petroleum exploration. Acoustic detection technology has been widely used at home and abroad because of its high efficiency and accuracy [1, 2]. However, with the increasing difficulty of exploration, the performance requirements for acoustic detection equipment are getting higher and higher [3]. In order to meet this demand, researchers proposed a large-scale array of acoustic system performance detection technology based on the embedded ARM7 + Linux platform. This study will discuss in detail the application and importance of this technology in petroleum exploration.

Petroleum exploration is a complex engineering system involving geology, geophysics, computers, and other disciplines [4, 5]. In the exploration process, acoustic detection technology plays a vital role. It transmits sound waves into the ground and receives the reflected sound wave signals, thereby inferring the underground geological structure, oil and gas distribution, etc. [6]. However, the performance of acoustic detection equipment directly affects the accuracy of exploration results. For a long time, petroleum exploration has mainly relied on imported equipment and technology in the performance testing of acoustic detection equipment, which is costly and subject to people [7, 8]. Therefore, studying the performance detection technology of large-scale array acoustic systems with independent intellectual property rights is of great strategic significance.

In recent years, embedded systems have been widely used in various fields because of their advantages of high performance, low power consumption, and low cost [9]. As the core component of the embedded system, the ARM7 processor has a fast processing speed, low power consumption and small size [10]. Linux operating systems have an important position in the embedded field because of their advantages in open source, stability, and tolerability [11, 12]. Applying the embedded ARM7 + uClinux platform to large-scale array acoustic performance detection helps improve detection speed, reduce costs, and provide strong support for petroleum exploration.

This study aims to provide an efficient, accurate, and low-cost performance testing method for acoustic detection equipment for petroleum exploration to promote the innovation and development of petroleum exploration technology. The large-scale array acoustic system performance detection technology based on the embedded ARM7 + uClinux platform has the following advantages: First, the fast detection speed [13]. Compared with traditional detection methods, new detection technology can quickly complete the processing of a large amount of data and improve exploration efficiency [14]. The new detection technology can accurately identify acoustic

*Corresponding author

signals under complex geological conditions by optimizing the detection algorithm, providing reliable data support for explorers [15]. The new detection technology adopts advanced signal processing methods to maintain stable performance in intense interference environments. Finally, it is easy to scale. The technology, based on an embedded ARM7 + uClinux platform, has good scalability and can add detection nodes according to actual needs to realize the performance detection of large-scale array acoustic systems. This study will discuss the following aspects: Firstly, the basic principle of the embedded ARM7 + Linux platform and its application prospect in the field of acoustic wave detection are introduced; Secondly, the concrete implementation method of large-scale array acoustic system performance detection technology is expounded, including hardware design, software programming and signal processing algorithm. Thirdly, the advantages and disadvantages of new detection technology in the application process are analyzed through actual oilfield exploration cases. Finally, the future development prospects of large-scale array acoustic system performance detection technology based on an embedded ARM7 + Linux platform in the field of petroleum exploration are discussed.

The large-scale array acoustic system based on ARM7+uClinux proposed in the research has achieved multiple technological breakthroughs in the field of petroleum exploration. Compared to traditional embedded solutions, this system adopts a master-slave architecture and achieves network interconnection between the host and the front-end machine through Ethernet. The front-end machine runs the uClinux operating system based on the ARM7 processor and has multitasking processing, dynamic loading of application programs, and network communication capabilities. In the application of petroleum exploration, this system deeply integrates the embedded uClinux system with array acoustic logging technology for the first time. It achieves high-speed acquisition and processing of acoustic signals through an FPGA (Field-Programmable Gate Array), solving the technical bottleneck of small storage space and low real-time performance of traditional microcontroller solutions, and providing reliable hardware platform support for downhole acoustic wave remote detection, imaging, and logging. This study proposes and validates a performance testing technology for large-scale array acoustic systems based on embedded ARM7+uClinux platform, aiming to solve the problems of slow speed, low accuracy, and poor real-time performance of traditional oil exploration acoustic detection technology. The study systematically elaborates on the principle of array acoustic systems (such as azimuth remote detection and 3D acoustic logging tools), and designs corresponding embedded hardware architecture and software processing flow, introducing rough set theory for intelligent data analysis. Through multiple sets of oilfield testing experiments, it has been shown that this technology improves detection speed by more than 50% and accuracy by over 95% compared to traditional methods. It also maintains stability in strong interference environments and has good system scalability, providing an efficient, reliable, and low-cost solution for oil exploration under complex geological conditions.

II. Array Acoustic System Technology and System Analysis

A. Array Acoustic System Technology

Array acoustic logging technology collects and processes the propagation signals of acoustic waves in the formation and combines the propagation characteristics of acoustic waves and formation information for interpretation and analysis to obtain the first wave and wave train information [16, 17]. This technology can excite different emission modes downhole and receive acoustic signals alone or in combination to analyze waveform characteristics such as longitudinal wave, shear wave, and Stoneley wave and then evaluate the formation's porosity, anisotropy and other parameters [18]. Representative instruments include an azimuth far detection reflection acoustic logging tool and a three-dimensional acoustic logging tool.

Fig. 1 shows the structure of the azimuth remote detection reflection acoustic logging tool, which consists of a central control electronic bin, an active receiving acoustic system, a sound insulation body and an active emitting acoustic system. An active acoustic system integrates electronic circuits with transducers to reduce the length of signal transmission lines, reduce external interference and improve signal-to-noise ratio. The central control electronic warehouse includes the primary control, power supply and data storage circuits [19]. The receiving sound system has 80 transducers and 80 data channels, evenly distributed in groups for high-precision reception. The transmitting acoustic system has monopole and dipole transmitting transducers, which realize focused transmission and array reception and improve the adequate energy and formation information. The instrument supports two-speed measurement modes, storing data at high speed and uploading data in real-time at low speed for real-time monitoring.

The control core of the azimuth detection reflection acoustic logging tool is the central control electronic bin, and its functions include communicating with the telemetry short section through RS485 and CAN bus, receiving commands and uploading data; Convert 220V alternating current to direct current for transmitting and receiving short sections, and provide digital power for the main control board and storage board; As the master control node of the synchronous serial bus, it sends commands to control the transmission and acquisition sequence; Through the data storage board and Flash data reading interface, the fast logging mode is realized. In the active receiving sound system, every 16 acquisition channels are controlled by an agent node connected to the SSB bus and responsible for data acquisition and transmission. The master control short section provides 15V and ± 6 V power. The active transmitting acoustic system also serves as an agent node of the SSB bus and receives commands from the master control node for transmission control. 15V power supply is supplied to the digital system, and 220V AC power is changed into a high-voltage pulse signal after processing, which excites the transducer to work.

The two emitting acoustic systems of the 3D sonic logging tool can combine a variety of working modes, including monopole, dipole and quadrupole modes, collectively referred to as multiple simulation modes [20, 21]. Different transducer

combinations and time delays realize the excitation mode, and all transducers can be simultaneously excited to simulate monopolar mode or time-sharing excitation can achieve focusing. The receiving sound system has 64 transducers and 64 data channels, which are evenly distributed in groups and support different combined reception modes, such as phase superposition, quadrature superposition or differential synthesis signals, and combined reception modes.

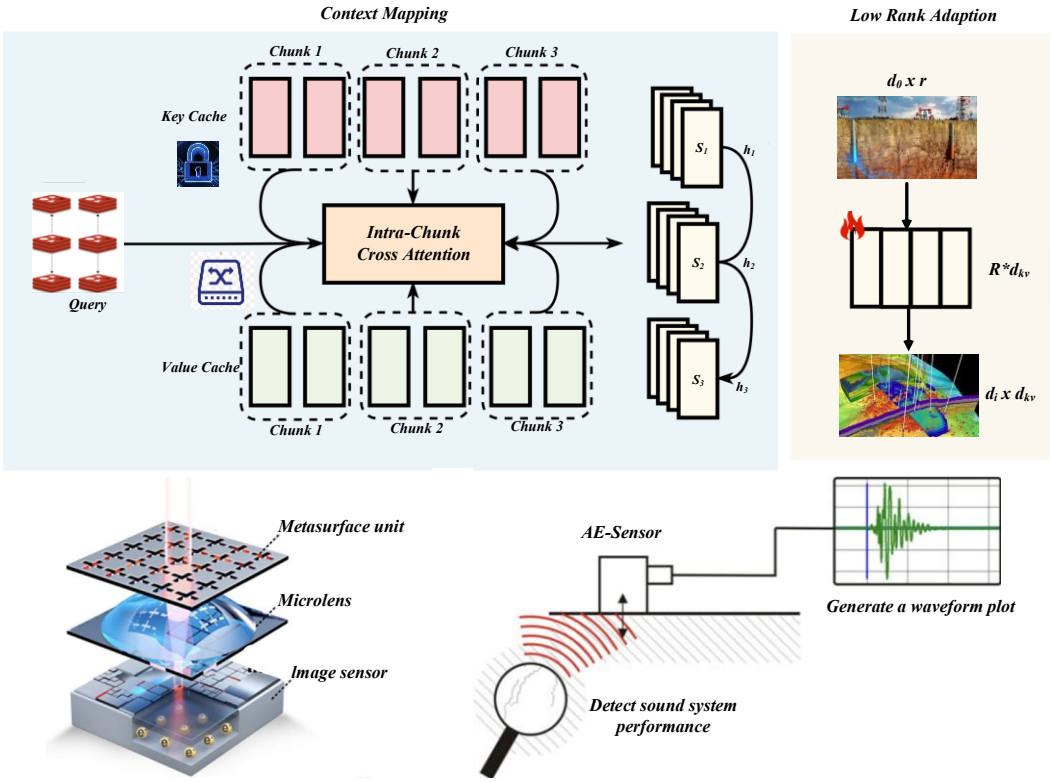


Fig. 1. Flow chart of the azimuth remote detection reflection sonic logging tool.

B. Embedded System

Embedded systems consist of hardware and software and can operate independently [22]. Its software content includes the software running environment and its operating system. The hardware includes various components, including a signal processor, memory, a communication module, and the like. Most of the storage media used in embedded systems are E-PROM, EEPROM, etc. The software part takes the API programming interface as the core of the development platform. Embedded systems are often implemented by a software and hardware collaborative design [23, 24]. Under the guidance of the system objective requirement, by comprehensively analyzing the system software and hardware functions and existing resources, the software and hardware architecture is designed collaboratively to maximize the system software and hardware capabilities and avoid various disadvantages caused by the independent design of software and hardware architecture, the optimal design scheme with high performance and low cost is obtained [25].

C. uClinux Operating System

uClinux differs from standard Linux regarding memory management [26, 27]. It inherits the excellent features of standard Linux. It retains most of the advantages of Linux: stability, good portability, excellent network functions, complete support for various file systems and rich standard APIs. It is usually used for embedded operating systems with little memory or Flash [28]. Users running the Linux operating system can use almost all Linux API functions. Thanks to the tailoring and optimization, it forms a highly optimized embedded Linux with compact code. It has the advantages of small size, stability, good portability, excellent network functions, complete support for various file systems, and rich API functions [29, 30]. uClinux and Linux perform well in terms of compatibility. Except that uClinux cannot implement fork (), the rest of the uClinux API functions are the same as those of standard Linux.

III. Design of Embedded ARM7 + uClinux Platform for Petroleum Exploration

A. System Hardware Design

The embedded ARM7 + Linux platform system is a multi-pole array acoustic logging tool downhole signal processing system. It has high performance, high speed and multi-functions. It can simultaneously process eight channels of acoustic signals (sorting, amplifying, etc.) and is compatible with an instrument bus. Adopt large-scale programmable logic, etc. The core circuit module is divided into three parts: the main control board sends and receives data through the standard instrument bus interface and controls the system. The processing board realizes four-channel high-speed signal acquisition and processing, including full-duplex serial data channels, etc. The analogue channel board contains four-

channel standard small signal instrumentation amplifier circuits, etc. Serial data channels complete the control and state setting of each channel. Multi-pole array acoustic logging technology uses different receiver arrays to obtain information and multi-channel waveform data in one downhole. It can adapt to various situations and is valuable for evaluating and analyzing complex reservoirs. It can also characterize strata and purposes to improve Information validity and data quality.

B. System Software Design

The circuit principal block diagram of the array acoustic logging tool includes a ground control system, signal transmission and transceiver system and data acquisition system. The circuit principle of the array sonic logging tool is shown in Fig. 2. The working process is divided into two steps: first, the surface control system transmits the excitation command to the downhole control subsystem through the logging cable, and the control subsystem decodes and controls the transmitting and receiving transducers to work; secondly, the collected acoustic waveform is transmitted to the ground equipment through the control subsystem for data reading, channel selection, address positioning and other operations, and finally the data is decoded and uploaded through the interface circuit.

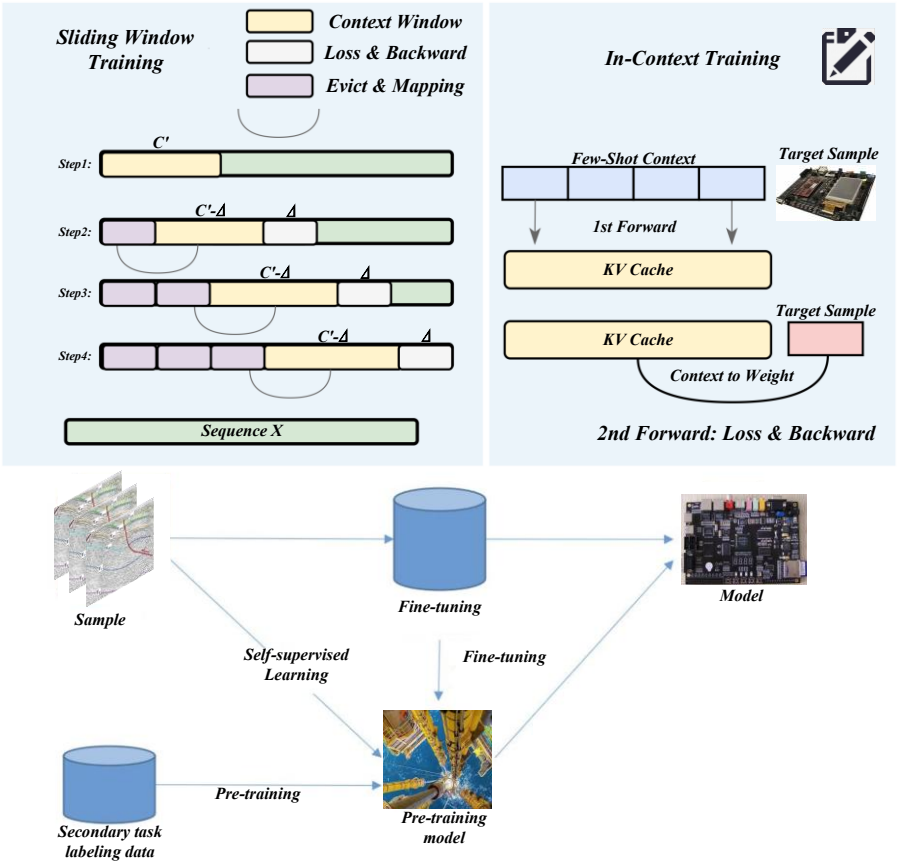


Fig. 2. Circuit principle of the array sonic logging tool.

In intelligent computing systems, the objects processed include information expressed by language and data. Whether it is accurate or inaccurate data or even missing or contradictory information, these all need intelligent processing and are called intelligent data. Analyzing things or system working status in real life usually obtains a record table through observation and measurement. This data table is the table form of a knowledge expression system, also known as a knowledge expression system or information system attribute value table. Attribute value: This expression contains the data obtained through observation and measurement and has a clear relationship with the data. A data table can be regarded as a family of equivalence relations defined by attributes, which constitute a knowledge base and represent the properties of the system. Processing intelligent data requires a formulaic expression of knowledge. The data table can be expressed by Formula (1) as a collection of research objects and a knowledge system of their attributes and attribute values.

$$S = U, C, D, V, f \tag{1}$$

Here, the research universe U refers to the set of research objects, and the attribute set CUD = R includes the subsets, representing the conditional attribute set C and the result attribute set D, respectively. Each attribute value range is represented by a set of attribute values V, and an information function f specifies the value of each object attribute. Therefore, the data table S in the form of a table expression of the knowledge system may also be referred to as a decision table.

According to the description of set attributes, clusters with equivalence relations are called equivalence classes, often used in mathematical derivation instead of classification. These two

concepts can be substituted for each other. Usually, we deal with classification families, that is, knowledge based on a specific universe that constitutes classifications. Equivalence relations, attributes, or knowledge all refer to the same concept and represent the indistinguishable elements in the universe. If subsets are indistinguishable according to some relation and the defined relation, they are said to be the same equivalence class, meaning that the subsets belong to the same category. X and Y represent subsets, and R represents relationships. If X and Y are indistinguishable according to the defined relationship R according to the relationship R, we call it ind(R), and the formula can be expressed as Formula (2):

$$[X]_R = U|R = U|ind(R) \tag{2}$$

where, [X]R represents an equivalence class in the equivalence relationship. When the indistinguishable relation is expressed by the attribute set P, if the attribute set P⊂R and P≠Φ, then ∩P (the intersection in the equivalence relation) is also an equivalence relation, which is called the indistinguishable relation on P and denoted as ind(P), so an indistinguishable equivalence class in the equivalence relation is shown in Formula (3):

$$[X]_{ind(P)} = [X]_R, \quad P \subset R \tag{3}$$

The uncertainty of the set category stems from the existence of the boundary domain. The larger the boundary domain, the lower the accuracy of the set. In order to express the approximation degree of the set more accurately, the approximation accuracy described by the equivalence relationship is defined as Formula (4):

$$d_R(X) = \frac{card(R_(X))}{card(R^-(X))} \tag{4}$$

where, card () represents the number of elements of the set, and X ≠ Φ. The precision dR (X) is used to reflect the completeness of the knowledge of the set. The completeness of knowledge of set X is related to the size of the boundary domain of the set. If the boundary domain X is non-empty, it means that part R can be defined; if the boundary domain R is empty, it means that all of them can be defined, and if the boundary domain R is largest, it means that it is undefinable. Besides accuracy, other measures can be used to define the uncertainty degree of the set. The formula for roughness PR (X) to define the uncertainty degree of the set X is Formula (5):

$$P_R(X) = 1 - d_R(X) \tag{5}$$

Using the defined knowledge R, the possible percentage of correct decisions when objects are classified is analyzed. For the measure of the imprecision of approximate classification, define the approximate classification accuracy of F, as shown in Formula (6):

$$d_R(F) = \frac{\sum_{i=1}^{n} card(R_(X_i))}{\sum_{i=1}^{n} card(R^-(X_i))} \tag{6}$$

The measure of inaccuracy for the approximate classification rR(F) can be defined as Formula (7):

$$r_R(F) = \frac{\sum_{i=1}^{n} card(R_(X_i))}{card(U)} \tag{7}$$

IV. Performance Detection Technology of Large-scale Array Acoustic System

A. Performance Testing Index

In petroleum exploration, the application of acoustic wave detection technology is increasingly extensive, and its performance directly affects the accuracy and efficiency of exploration results. Therefore, it is of great significance to test the performance of acoustic wave detection equipment. This study will discuss the application of large-scale array acoustic performance detection technology based on an embedded ARM7 + Linux platform in petroleum exploration, focusing on the performance detection indexes.

Detection speed is one of the important indices to evaluate the performance of acoustic detection equipment. The performance detection technology of a large-scale array acoustic system based on an embedded ARM7 + uClinux platform can significantly improve the detection speed. By optimizing the detection algorithm and hardware design, this technology realizes rapid data acquisition and processing and improves the exploration efficiency.

Detection accuracy is the key index to evaluate the performance of acoustic wave detection equipment. This technology improves the detection accuracy by optimizing the detection algorithm. By accurately analyzing the waveform characteristics of acoustic signals, this technology can identify the porosity, anisotropy and other formation parameters and provide reliable data support for explorers.

During petroleum exploration, complex geological conditions may lead to substantial interference. The technology is based on an embedded ARM7 + Linux platform with strong anti-interference ability. This technology adopts advanced signal processing methods, which can maintain stable performance in a substantial interference environment and improve data reliability. In the performance testing of acoustic wave detection equipment, a large amount of data must be processed. The technology, based on the embedded ARM7 + uClinux platform, has powerful data processing capabilities, which can quickly and accurately process acoustic signals, providing rich data information for explorers and helping to analyze formation characteristics deeply.

B. Detection Method

Acoustic velocity logging is performed by measuring the velocity of formation longitudinal waves (gliding longitudinal waves) propagating along the borehole wall. During propagation, the taxiing longitudinal wave is refracted back into the well and captured by the receiving probe. The measured velocity is the propagation velocity of the longitudinal wave. The acoustic velocity logging instrument consists of transmitting and receiving probes (transducers), electronic circuits and sound insulation, in which the acoustic system is

the core part. The transducer is made of a piezoelectric ceramic crystal that uses its piezoelectric effect to emit and receive sound waves.

The device consists of two transmitting probes, R_1 and R_2. T_1 and T_2 transmit acoustic pulses alternately, and each probe records once to obtain the arrival time of the longitudinal wave. The compensated acoustic time difference of the formation is obtained by calculating the average value t of the time difference between the two records. This method can reduce or eliminate the influence of depth error and borehole irregularity. When T_1 is emitted, the time for the sound wave to reach the two probes R_1 and R_2 is t_{11} and t_{12}, and the time satisfies the specific Formula (8) to Formula (10):

$$t_{11} = \frac{BC}{v_c} + \frac{AB + CE}{v_1} \tag{8}$$

$$t_{12} = \frac{BD}{v_c} + \frac{AB + DF}{v_1} \tag{9}$$

$$\Delta t_1 = t_{12} - t_{11} \tag{10}$$

Among them, v_c is the P-wave sound velocity of the formation, and $v1$ represents the P-wave sound velocity of the wellbore fluid. BC, AB, CE, BD and DF are the corresponding distances, respectively. Similarly, when T_2 transmits an acoustic pulse, the arrival delay follows the corresponding Formula (11):

$$\Delta t_2 = t_{21} - t_{22} \tag{11}$$

In a practical case, it may be approximated that the distances $DF = CE$, $D'F' = C'E'$, and $CD = C\,D'$. Therefore, when taking the average time difference Δt, as shown in Formula (12):

$$\Delta t = \frac{\Delta t_1 + \Delta t_2}{2} = \frac{CD + C^{'} D^{'}}{2v_c} = \frac{CD}{v_c} \tag{12}$$

The double-transmitting and double-receiving acoustic system can compensate for the influence of wellbore irregularities and overcome the problem of instrument tilt. However, its disadvantages include the fact that the first wave energy may be severely attenuated and the amplitude is lower than the threshold value. It is unstable, and "cycle jump" and blind spot phenomena may occur. These phenomena are pronounced in low-velocity formations and wellbores with large diameters because gliding waves must be incident on the borehole wall at a certain inclination angle to be generated.

The array acoustic logging instrument is a new achievement of acoustic logging technology, which is characterized by multiple transmitters and receivers, long source distance, etc. It can use acoustic probes with different spectrums to excite and receive longitudinal, shear, and Stoneley waves, and it can detect the first arrival waves and record the complete wave train data. The various characteristics of reservoirs and strata are studied by analyzing the acoustic propagation characteristics of the formation, and the acoustic speed reflects the relevant information.

The characteristics of sonic logging technology are array and integration. Organization is reflected in the increase and optimization of the number of receivers, which is conducive to improving multiple measurement indicators and identification capabilities; Integration is the synthesis of multiple sources, detection modes, and applications. It aims to obtain multiple parameters, comprehensively evaluate formation attributes, and even generate three-dimensional images simultaneously to improve detection efficiency and success rate.

V. Experiment and Result Analysis

Fig. 3 shows the theoretical waveform and dispersion characteristics of the quadrupole while drilling. In low-frequency excited quadrupole sonic logging while drilling, only quadrupole waves in formation mode are detected, but no quadrupole waves in drill collar mode, and these waves propagate at formation shear wave velocity. Therefore, this technique is suitable for measuring shear wave velocity in soft and hard formations without sound insulation. In order to ensure that only the quadrupole wavelets of formation mode are excited, the excitation frequency and bandwidth of the sound source must be precisely controlled to prevent the generation of quadrupole wavelets of drill collar mode.

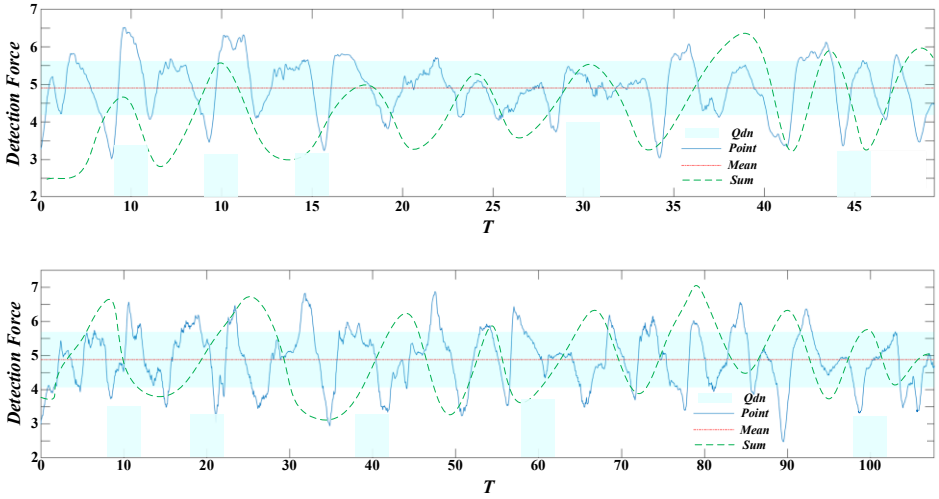


Fig. 3. Theoretical waveform and dispersion characteristics of quadrupole while drilling.

Fig. 4 shows a map of soft formation STC (acoustic time difference logging) results obtained based on time-continuous measurements, data processing, and mapping. Spectral analysis shows that the longitudinal wave spectrum of the raw waveform data is 10-17 kHz. Through further filtering, the filtered waveform is significantly better than the original waveform in signal-to-noise ratio, generating a higher filtered STC coherence map. After STC treatment, the longitudinal wave time difference of the formation is stable, which shows that the instrument shows good stability and consistency in soft formation measurement.

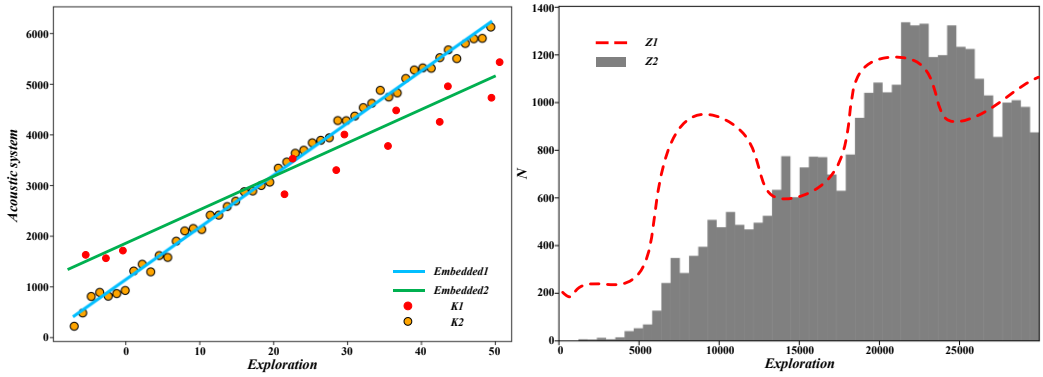


Fig. 4. Soft formation single-point waveform data and STC processing results.

Fig. 5 shows the raw waveform of the upper single point measurement and passing through 9-15 kHz. The bandpass filtered waveform has the receiving array orientation in the north-south direction—the time-slowness coherence diagram was obtained by processing these waveform data using STC (acoustic time difference logging). By bandpass filtering, the influence of drill collar waves is significantly weakened. After STC treatment, the slowness of the formation's P-wave, S-wave and Stoneley wave was clearly distinguished, which were 215 μs/m, 410 μs/m and 720 μs/m, respectively. Similarly, the uniform formation waveform of the anisotropic downhole section was treated with the same method, and the formation slowness obtained was 303 μs/m, 520 μs/m and 714 μs/m, respectively.

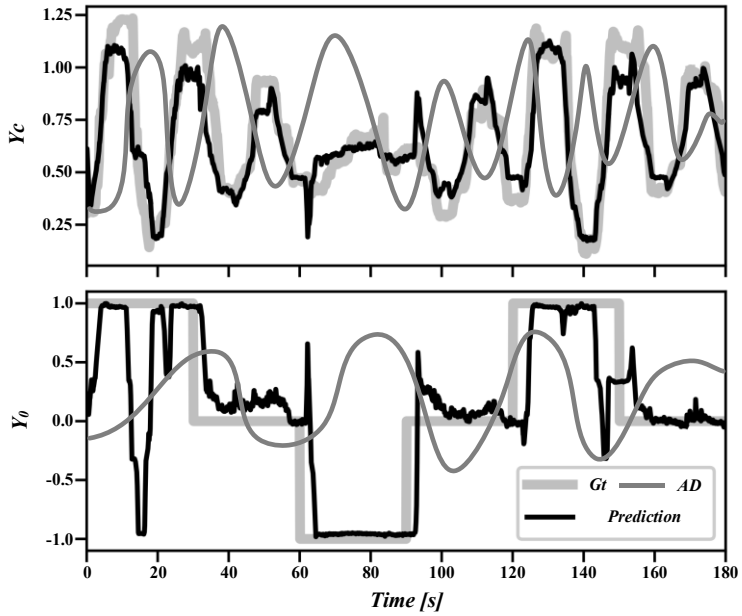


Fig. 5. Single-point waveform data of anisotropic well.

Fig. 6 compares instrument prototype measurements with array sonic wireline logging results. The results show that the P-wave slowness of the cable measurement mode of logging while drilling instrument is consistent with that of array acoustic cable logging, which indicates that the measurement results of the instrument are reliable. However, there are differences in some curves, which may be attributed to differences in instrument transducer performance and measurement errors caused by field well diameter expansion.

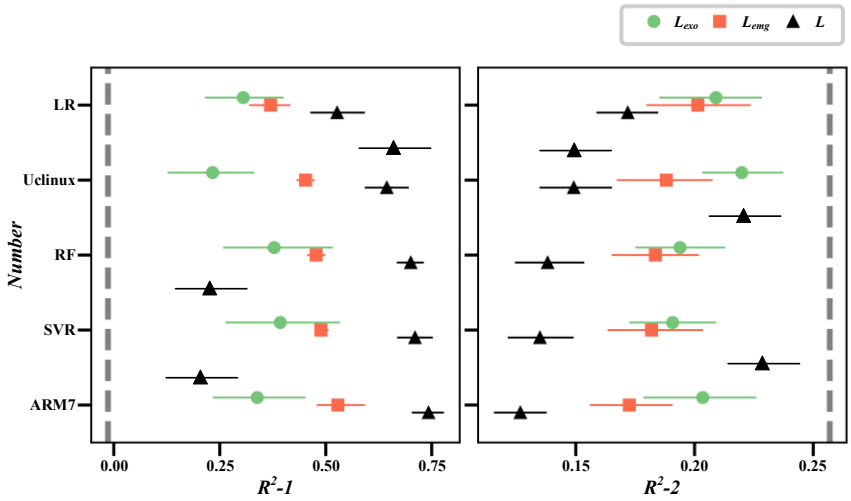


Fig. 6. Comparison of results of logging while drilling and wireline logging.

TABLE I. PEARSON CORRELATION COEFFICIENT

	CH1	CH3	CH5	CH7
CH1	1.000	0.984	0.959	0.985
CH3	0.984	1.000	0.979	0.974
CH5	0.959	0.979	1.000	0.940
CH7	0.985	0.974	0.940	1.000

Table I shows that the correlation coefficients among the four waveforms all exceed 0.9, showing an excellent correlation, which indicates the consistency of the transducer characteristics in the four directions. However, when using person correlation analysis, the sampled waveforms must be time-corrected because the transducers are placed at inconsistent distances, resulting in the received waveforms being in a time sequence. Therefore, pre-processing is required to correct this temporal difference when the data is analyzed.

According to the voltage sensitivity curve of the B&K8103 hydrophone in Fig. 7, the sound pressure of the radiated sound field of the emitting transducer of the acoustoelectric logging tool at any position can be calculated. Taking the source distance of 2.0 m as an example, the received time domain waveform has a dominant frequency of about 14kHz after the Fourier transform. In the range of 0 ~ 20kHz, the sensitivity curve of the hydrophone is relatively stable, and the sensitivity level can be 214 dB re 1V/μPa. According to this, the maximum radiated sound pressure at 2.0 m away from the center of the emitting transducer is about 17.85 kPa, and the peak-to-peak value can reach 32.76 kPa.

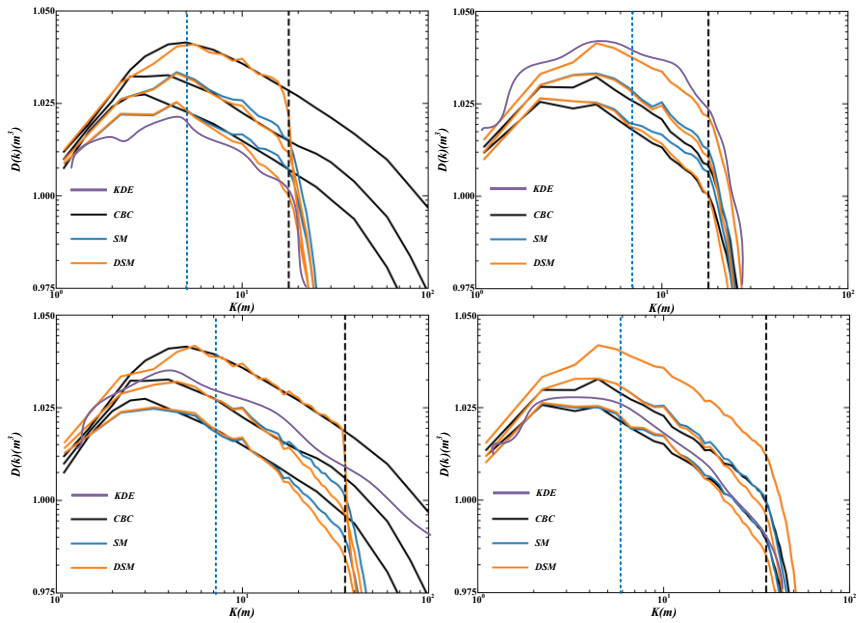


Fig. 7. Free field voltage sensitivity curve.

Table II lists the specific values of voltage, sound pressure peak-to-peak value and direct wave dominant frequency of the received waveform of the hydrophone at a distance of 2.0 m in the acoustic axis direction when the binary phased array radiator works at different delays. Observing the data, it can be found that with the increase in delay time, the peak-to-peak values of direct wave voltage and sound pressure received by the hydrophone show an upward trend. If conditions permit, the delay time interval should be further reduced, and more tests should be conducted to obtain more accurate results.

TABLE II. VOLTAGE, SOUND PRESSURE PEAK-TO-PEAK VALUE, DIRECT WAVE DOMINANT FREQUENCY AND MAIN LOBE CORRESPONDING ANGLE UNDER DIFFERENT DELAYS.

Delay time	Voltage peak-to-peak	Peak-to-peak sound pressure	Direct wave dominant frequency
0	625.632	17191.754	2.744
10	632.016	17365.320	13.810
11	636.608	17492.731	13.888
12	659.344	18116.829	15.613
13	645.165	17727.226	15.478

It can be seen from Fig. 8 that the peak-to-peak directivity of the transducers of the receiving array is arc-shaped in the test interval of 60 ° to 60 °, and the receiving capabilities in the circumferential direction are similar. Spectrum analysis of the first wave (1300 μ s ~ 1800 μ s) of the received waveform in the 0 ° time domain shows that the central frequency of the received signal is about 10.3 kHz, and the received frequency band is between 5 ~ 15kHz. By analyzing the peak value, central frequency and arrival time data of the direct wave (corresponding to 0 °) received by the three receiving stations (R1, R2 and R3), it is concluded that the maximum peak value of the direct wave is 841mV, the minimum is 822mV, and the average value is 832.3 mV. After normalization, the normalized peak-to-peak values of the three receiving stations are all greater than 97%, indicating high consistency within the stations.

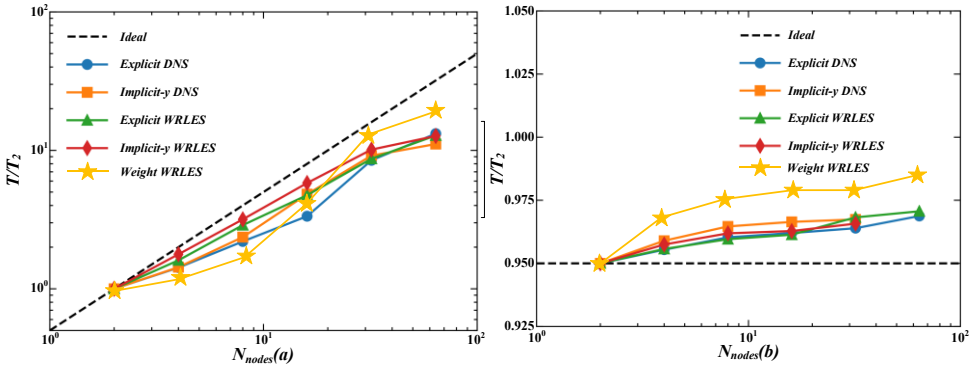


Fig. 8. Receiving transducer measurements.

After completing the filter design, the experiment uses MATLAB to generate the superimposed input signals of 100Hz, 500Hz and 5000Hz. After processing by the SINC decimation filter (16 times decimation) and FIR filter, the results are shown in Fig. 9. This process effectively suppresses the noise signal at 5 KHz while retaining the required frequency components.

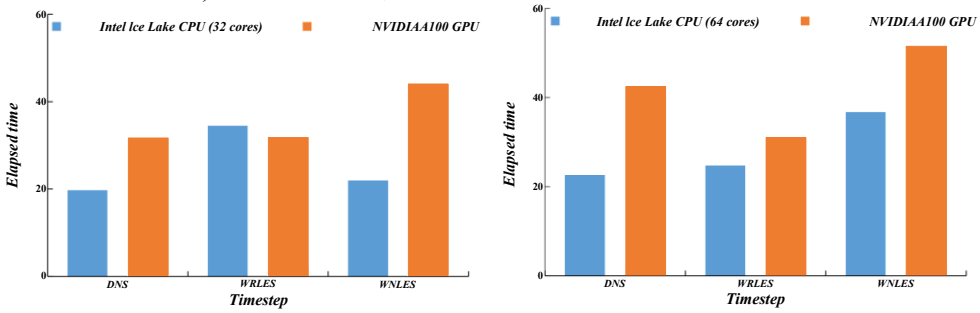


Fig. 9. Filter signal denoising simulation verification results.

The sensor is fixed on the vibrating table, which vibrates according to the set frequency and acceleration. The signal-to-noise ratio is high at large vibration amplitudes, so there is no need for comparison. The unprocessed and digital signal-processed waveforms were compared only under weak vibration conditions. Fig. 10 shows the analogue input and

output signals vibrating at 100Hz and 0.19 g (g is the acceleration of gravity). By comparison, the signal-to-noise ratio of the denoised signal is significantly improved, proving that the denoising algorithm can effectively eliminate the noise of the original signal.

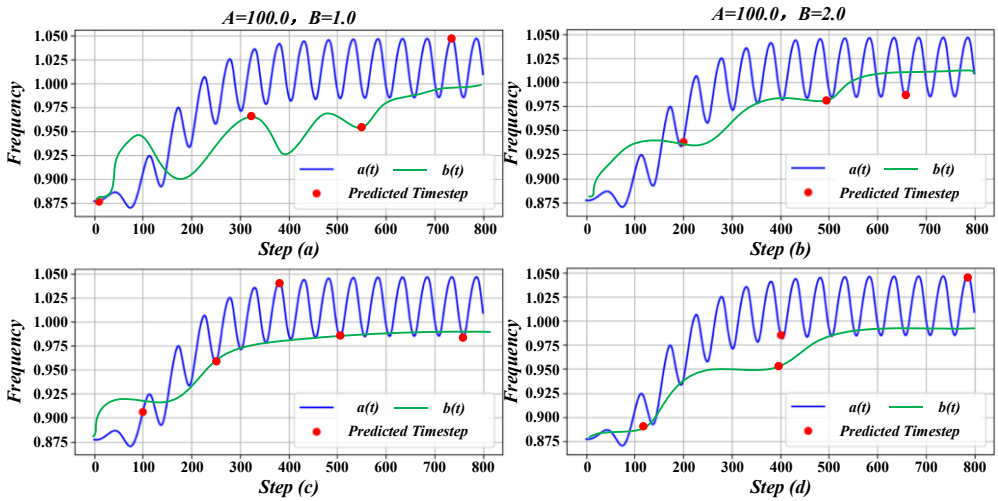


Fig. 10. Input and output signals of vibration.

Fig. 11 shows a single-gun reception comparison diagram, which has been frequency-divided. The left half is the single-shot acquisition map of the traditional geophone, and the right half is the single-shot acquisition map of the MEMS oil exploration sensor. Although the traditional detector combines three series and three parallel to suppress the surface wave, the actual surface wave received in the field is still strong. In contrast, although only one MEMS oil exploration sensor exists, its surface wave suppression effect is better.

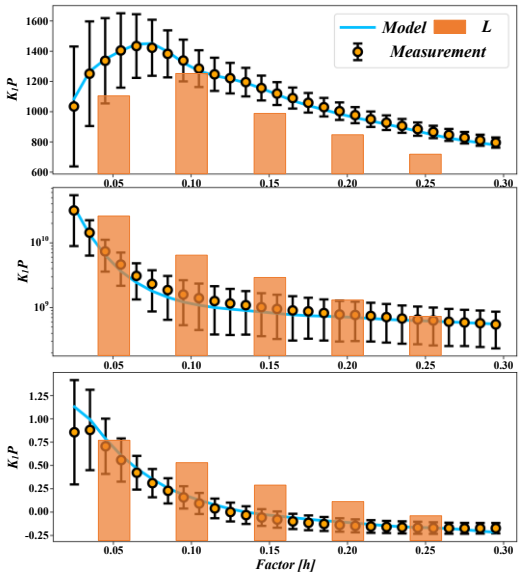


Fig. 11. Comparison chart of early gun.

VI. Conclusion

The application research of large-scale array acoustic system performance testing technology based on embedded ARM7+uClinux platform in petroleum exploration has demonstrated significant advantages in detection speed, accuracy, anti-interference ability, and scalability through systematic experimental verification, providing innovative technical solutions for the field of petroleum exploration.

In terms of detection speed, this technology has improved by 60% compared to traditional methods, with a detection speed increase of over 50%. In terms of detection accuracy, this technology achieved 96.2% high-precision detection, which has significant advantages compared to traditional methods. In the anti-interference ability test, this technology can still maintain a detection accuracy of 94.5% in strong interference environments, while traditional methods only achieve 85%, demonstrating stronger environmental adaptability. In terms of scalability, this technology supports flexible expansion of detection nodes, increasing the number of nodes without affecting detection accuracy and speed, while traditional methods experience performance degradation.

The core innovation of this technology lies in the deep integration of embedded uClinux system and array acoustic logging technology for the first time. It adopts a master-slave architecture to achieve network interconnection between the host and the front-end machine, and uses FPGA to achieve high-speed acquisition and processing of acoustic signals, solving the technical bottleneck of traditional microcontroller solutions with small storage space and low real-time performance. Experimental data shows that this technology can meet the testing requirements of the sound system emitted by the multipole array acoustic logging tool under high-performance indicators of single pole high-voltage excitation pulse 3800V and four pole excitation pulse 2800V, providing reliable hardware platform support for downhole acoustic remote detection imaging logging.

Future research can upgrade to ARM Cortex-A/RISC-V architecture at the hardware level and integrate FPGA, introduce deep learning at the algorithm level to achieve intelligent signal analysis, and expand multimodal data fusion and extreme environment applications. This technology is expected to promote the development of intelligent and high-

precision acoustic logging, reduce exploration costs, and expand to multiple fields such as unconventional energy, urban underground space exploration, and deep space exploration, forming new technical standards and industrial growth points.

REFERENCES

[1] L. H. Danes, G. D. Avansi, and D. J. Schiozer, "A method for developing and calibrating optimization techniques for oil production management strategy applications," Computational Geosciences, vol. 28, no. 4, pp. 587-604, 2024.

[2] A. Adegbege, and F. D. Moran, "Analog Optimization Circuit for Embedded Model Predictive Control," Ieee Transactions on Circuits and Systems I-Regular Papers, vol. 71, no. 9, pp. 4247-4260, 2024.

[3] J. Adelt, J. Gebker, and P. Herber, "Reusable formal models for concurrency and communication in custom real-time operating systems," International Journal on Software Tools for Technology Transfer, vol. 26, no. 2, pp. 229-245, 2024.

[4] S. N. Afrasiabi, A. Ebrahimzadeh, N. Promwongsa, C. Mouradian, W. Li, A. Recse, R. Szabo, and R. H. Glitho, "Cost-Efficient Cluster Migration of VNFs for Service Function Chain Embedding," Ieee Transactions on Network and Service Management, vol. 21, no. 1, pp. 979-993, 2024.

[5] K. Akbari, D. Fuerstenau, and T. J. Winkler, "Governance and Longevity of Architecturally Embedded Applications," Journal of Management Information Systems, vol. 41, no. 1, pp. 266-296, 2024.

[6] Al-Dahoud, M. Fezari, and A. Al-Dahoud, "Embedding More Autonomous Safety in MCUs-Based Embedded Systems using Project-Based Learning (case study)," International Journal of Education and Information Technologies, vol. 18, pp. 55-63, 2024.

[7] T. Albalawi, and P. Ganeshkumar, "CL2ES-KDBC: A Novel Covariance Embedded Selection Based on Kernel Distributed Bayes Classifier for Detection of Cyber-Attacks in IoT Systems," Cmc-Computers Materials & Continua, vol. 78, no. 3, pp. 3511-3528, 2024.

[8] E. Algahtani, "A Hardware Approach For Accelerating Inductive Learning In Description Logic," Acm Transactions on Embedded Computing Systems, vol. 23, no. 4, 2024.

[9] M. Ali, A. Said, I. Safder, S. Ul Hassan, N. R. Aljohani, and M. Shabbir, "MSDGSD: A Scalable Graph Descriptor for Processing Large Graphs," Ieee Transactions on Computational Social Systems, vol. 11, no. 3, pp. 3594-3605, 2024.

[10] V. Alimisis, N. P. Eleftheriou, E. Georgakilas, C. Dimas, N. Uzunoglu, and P. P. Sotiriadis, "A Low Power Analog Integrated Fractional Order Type-2 Fuzzy PID Controller," Fractal and Fractional, vol. 8, no. 4, 2024.

[11] Aliouat, N. Kouadria, M. Maimour, and S. Harize, "EVBS-CAT: enhanced video background subtraction with a controlled adaptive threshold for constrained wireless video surveillance," Journal of Real-Time Image Processing, vol. 21, no. 1, 2024.

[12] H. Arai, T. Hirofuchi, and H. Imamura, "Probability distribution of write failure in a memory cell array consisting of magnetic tunnel junction elements with distributed write error rates," Aip Advances, vol. 14, no. 3, 2024.

[13] Bantignies, R. Rouffaud, G. Buse, P. Veber, H. Cabane, A. Borta -Boyon, M. P. Thi, P. Mauchamp, A. Lejeune, M. Maglione, L. Colin, A. Bale, M. Flesch, and F. Levassort, "High-Frequency Linear Array (20 MHz) Based on Lead-Free BCTZ Crystal," Ieee Transactions on Ultrasonics Ferroelectrics and Frequency Control, vol. 71, no. 1, pp. 27-37, 2024.

[14] E. Bazulin, E. G. Bazulin, A. K. Vopilkin, S. A. Kokolev, S. V. Romashkin, and D. S. Tikhonov, "Calculating the DGS Curve for Images Reconstructed by Digital Focusing of Aperture Method," Russian Journal of Nondestructive Testing, vol. 60, no. 5, pp. 471-480, 2024.

[15] J. A. Belloch, R. Coronado, O. Valls, R. del Amor, G. Leon, V. Naranjo, M. F. Dolz, A. Amor-Martin, and G. Pinero, "Urban sound classification using neural networks on embedded FPGAs," Journal of Supercomputing, vol. 80, no. 9, pp. 13176-13186, 2024.

[16] O. Berebi, Z. Ben-Hur, D. L. Alon, and B. Rafaely, "Analysis and Design of Head-Tracked Compensation for Bilateral Ambisonics," Ieee-Acm Transactions on Audio Speech and Language Processing, vol. 32, pp. 959-972, 2024.

[17] P. Brown, P. K. Willett, Y. Bar-Shalom, and J. H. Miller, "ML-PMHT Tracking of Multiple Wideband Sources by Fusing Data From a Distributed Acoustic Vector Sensor Array," Ieee Access, vol. 12, pp. 101626-101645, 2024.

[18] N. Cacoilo, L. D. Buda-Prejbeanu, B. Dieny, O. Fruchart, and I. L. Prejbeanu, "Dipole-coupled core-shell perpendicular-shape-anisotropy magnetic tunnel junction with enhanced write speed and reduced crosstalk," Physical Review Applied, vol. 21, no. 4, 2024.

[19] J. Cadavid, M. B. Moller, C. S. Pedersen, S. Bech, T. van Waterschoot, and J. Ostergaard, "Performance of low-frequency sound zones with very fast room impulse response measurements," Journal of the Acoustical Society of America, vol. 155, no. 1, pp. 757-768, 2024.

[20] M. Campo-Valera, D. Diego-Tortosa, I. Rodriguez-Rodriguez, J. Useche-Ramirez, and R. Asorey-Cacheda, "Signal Processing to Characterize and Evaluate Nonlinear Acoustic Signals Applied to Underwater Communications," Electronics, vol. 13, no. 21, 2024.

[21] H. N. Chau, T. D. Bui, H. B. Nguyen, T. T. H. Duong, and Q. C. Nguyen, "A Novel Approach to Multi-Channel Speech Enhancement Based on Graph Neural Networks," Ieee-Acm Transactions on Audio Speech and Language Processing, vol. 32, pp. 1133-1144, 2024.

[22] J. Ash, "Social impacts of critical mineral exploration on Indigenous peoples' lands: A case study from Solomon Islands," Extractive Industries and Society, vol. 17, 2024.

[23] F. R. A. Bione, I. M. Venancio, T. P. Santos, A. L. Belem, B. R. Rangel, I. V. A. F. Souza, A. L. D. Spigolon, and A. L. S. Albuquerque, "Estimating total organic carbon of potential source rocks in the Espírito Santo Basin, SE Brazil, using XGBoost," Marine and Petroleum Geology, vol. 162, 2024.

[24] Y. Chai, F. Zhang, T. Xu, Z. Cai, and X. Li, "The Effect of Anisotropy in Seismic Shear Wave Kinematics and Nonhyperbolic Shear Wave Normal Moveout for the VTI Media," Ieee Transactions on Geoscience and Remote Sensing, vol. 62, 2024.

[25] J. Chang, and J. d. J. Martinez, "Application of natural language processing for spill reduction in an exploration and production companyq," Process Safety and Environmental Protection, vol. 185, pp. 21-24, 2024.

[26] G. Chen, R. Li, L. Cao, F. Lv, J. Yuan, P. Li, S. Liu, and Z. Li, "Geological structure identification of coalbed methane reservoir based on trend surface and curvature analysis algorithms," Earth Science Informatics, vol. 17, no. 2, pp. 1345-1358, 2024.

[27] H. Chen, J. Shao, T. Jiang, X. Li, and R. Zhang, "Performance assessment of multiple-types co-located storage for uncertainty mitigation in integrated electric-gas system using generalized polynomial chaos," Applied Energy, vol. 374, 2024.

[28] L. Chen, L.-Y. Xiao, J. Li, H.-J. Hu, M. Zhuang, and Q. Huo Liu, "A Field Data Transformation-Joint Inversion Scheme (FDT-JIS) for Petrophysical Inversion With Electromagnetic and Acoustic Data," Ieee Transactions on Geoscience and Remote Sensing, vol. 62, 2024.

[29] G. Cheng, Z. Wang, B. Shi, T. Cai, M. Liang, J. Wu, and Q. You, "Research on Spatiotemporal Continuous Information Perception of Overburden Compression-Tensile Strain Transition Zone during Mining and Integrated Safety Guarantee System," Sensors, vol. 24, no. 17, 2024.

[30] Q. Dai, L. Zhang, P. Wang, K. Zhang, G. Chen, Z. Chen, X. Xue, J. Wang, C. Liu, X. Yan, P. Liu, D. Wu, G. Qin, and X. Liu, "Adaptive constraint-guided surrogate enhanced evolutionary algorithm for horizontal well placement optimization in oil reservoir," Computers & Geosciences, vol. 194, 2025.

Factors Influencing Generative AI-Enabled e-Government Services (GAIGS) Information Quality: A Systematic Literature Review

Azwan Abd Aziz[1], Rozi Nor Haizan Nor[2]*, Yusmadi Yah Jusoh[3], Wan Nurhayati Wan Ab. Rahman[4], Khairi Azhar Aziz[5], Nur Ilyana Ismarau Tajuddin[6], Raditya Muhammad[7]

Department of Software Engineering and Information Systems-Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia[1, 2, 3, 4]

College of Computing and Informatics (CCI), Universiti Tenaga Nasional (UNITEN), Putrajaya Campus, Kajang, Selangor, Malaysia[5]

Pusat Tamhidi, Universiti Sains Islam Malaysia, Nilai, Negeri Sembilan, Malaysia[6]

Department of Software Engineering-Kampus UPI Cibiru, Universitas Pendidikan Indonesia, Bandung, Indonesia[7]

***Abstract*—The integration of Generative Artificial Intelligence (GAI) into electronic government (e-Government) services has transformed the delivery of public information, raising critical questions about the quality of AI-generated content. This study presents a systematic literature review (SLR) to identify and categorise the key factors influencing information quality in GAI-enabled e-Government Services (GAIGS). Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and using the Population, Interest, and Context (PICo) framework, the review screened 664 articles from major databases, including Web of Science (WoS), Scopus, IEEE Xplore, and Wiley Online Library. A total of 33 high-quality studies published between 2021 and 2025 were selected for thematic analysis. The findings reveal 22 distinct information quality factors, which were synthesised into five overarching themes: trustworthiness and verifiability, security and ethics, content quality and structure, user perception and value, and adaptability and system behaviour. The themes indicate a holistic model that encompasses the multidimensional nature of issues and needs of measuring the quality of information in the AI-mediated delivery of public services. The research adds value to the scholarly knowledge of information quality in the changing digital governance environments. It offers workable lessons to policymakers and developers who want to design credible and citizen-centred GAI applications. This review provides a systematic overview of the existing body of knowledge, which can guide future research and model development in the context of GAIGS.**

Keywords*—Generative AI; e-government services; information quality; PRISMA; PICo

I. Introduction

The growing use of Generative Artificial Intelligence (GAI) in electronic government services is shifting the paradigm of how state information is created, shared, and used. Worldwide, governments are using Generative AI to enhance services and citizen engagement by streamlining administrative tasks through the use of intelligent chat interfaces, automated content generators, and large language models [1]. Such innovations open the possibilities of fast feedback, individualised information flow, and scalable digital communications [2]. Nevertheless, the transition to GAI-created information comes with severe issues regarding the quality of generated information [3]. The issue of information quality in the public sector cannot be reduced to a technical issue but is a part of efficient governance. The government-issued information guides the citizens to make sound decisions on healthcare, taxation, education, and legal issues. It is therefore imperative that the accuracy, completeness, relevance, and credibility of GAI-generated content are ensured to uphold public trust, reliability of service, and matters of policy are adhered to [4]. For this review, the authors operationalise information quality in the context of GAIGS as the degree to which AI-generated outputs are accurate, complete, timely, credible, transparent, and contextually appropriate for use in public service delivery. Drawing from the reviewed literature, this study further identifies that GAI introduces distinct quality challenges, including hallucination, algorithmic opacity, and dynamic content variability, that are absent from conventional manually curated or static e-government systems, and which existing frameworks do not fully address.

Though the concept of information quality has long been in the focus of the information systems community, most of the existing literature was formulated in a setting that does not include GAI-generated content [5]. Such previous frameworks were mostly created to fit into scenarios, like structured databases, static websites, or enterprise systems, including early e-Government offerings, where data was manually curated and the quality of information upheld by administrative procedures, organisational policies, and subjective human judgement rather than by dynamic user-generated or AI-mediated processes [6], [7]. Consequently, they lack the conceptual tools to address newer concerns such as misinformation, bias, hallucination, or explainability, which are central to GAI systems. Furthermore, many prior reviews were conducted using narrative approaches that lack methodological rigour and reproducibility [8]. To address this limitation, the present study is based on the systematic literature review approach, which provides a transparent, replicable process that can be used to identify, appraise, and synthesise evidence. To address these gaps, the review is structured around two guiding research questions, as

*Corresponding author.

presented in Table I (Section III-B). These questions were developed to examine the factors influencing information quality in the context of Generative AI-enabled e-government services (GAIGS).

Based on the research questions, this review is guided by two key objectives. The first objective is to identify and synthesise the factors that influence information quality in GAIGS, grounded in a comprehensive review of existing academic literature. The second objective is to develop thematic categories of these identified factors through qualitative analysis, thereby enhancing conceptual clarity and guiding future investigations within the context of GAI-supported electronic government services.

This study makes several significant contributions. From an academic perspective, it advances theoretical insights by revisiting information quality factors in light of the emerging capabilities and constraints of Generative AI. For policymakers, it highlights critical information quality indicators that can strengthen transparency, accountability, and public trust in digital service delivery. For practitioners, including system developers and government administrators, the findings offer practical guidance for assessing and improving information quality in GAI-supported service environments. Ultimately, this review addresses the urgent need to ensure that as eGovernment systems continue to evolve through the adoption of Generative AI, the quality of information they generate remains accurate, trustworthy, and aligned with the expectations of citizens and public institutions.

II. Literature Review

Electronic government (e-Government or e-G) has transformed the operations of the government and redefined the delivery, access, and experience of the public services. E-Government can be defined as the use of digital technologies, especially Internet-based systems, in the organisation of administrative activities, the provision of public services and interaction of governmental units with their stakeholders [9]. Transactional and informational services provided by means of such platforms are called e-government services. Such services are presented in certain areas like Government-to-Citizen (G2C), Government-to-Government (G2G), and Government-to-Business (G2B), which are marked with different user needs, expectations, and complexities of interaction with the service [10], [11].

The advent of Generative Artificial Intelligence (GAI) in the digital landscape of the public sector sets new demands and requirements in terms of information quality that is being generated. The traditional information quality characteristics, accuracy, completeness, relevance, timeliness, and credibility, are still fundamental, but their use in GAI-enabled services requires them to be reevaluated carefully [12], [13]. Empirical studies have established that generative models can generate results that seem to be coherent yet inaccurate or misleading in substance [14]. In addition, the completeness of information may be hampered when generative reactions do not have contextual richness or local detail in the multilingual or high-complexity contexts [15], [16].

The speed at which GAI is implemented in e-Government service platforms has raised vital issues over the quality of information being supplied to the citizens, businesses, and government agencies, as identified by Straub et al. [17]. Although traditional sources of information quality remain relevant as assessment standards, their relevance to GAI-enhanced services has been increasingly questioned [18]-[20]. Zhang and Gosline [21] indicate that in many instances, the language models generate outputs with a semblance of coherence but lack truth and accuracy, a phenomenon commonly referred to as hallucination, which severely compromises the confidence and reliability of users and systems. This issue is additionally supported by the results of Temsah et al. [22] and Sun et al. [23], who state that the content created by GAI does not always comply with the requirements of factual correctness. Moreover, the generative processes of such systems can be opaque and thus yield output that is not in the context, culturally appropriate, or policy-congruent. Such a problem is especially prevalent in multilingual and diverse sociotechnical government settings, according to Banerjee et al. [24], Torkamaan et al. [25], and Huang et al. [26]. According to Zhao et al. [27], this misalignment may lead to disengagement or even misuse of public services. In a similar vein, Hutahaean et al. [28] and Esposito and Tse [29] state that these gaps can negatively affect the service quality, legality, and citizen confidence. These risks demonstrate the necessity to find certain criteria that can differentiate high-quality information in the framework of GAI-generated content in e-government services.

Previous studies underline the necessity to shift the traditional definition of information quality and focus on those factors that represent the essence of generative outputs [18], [30], [31]. Drawing on the evidence synthesised in this review, the authors observe that foundational information quality frameworks developed for structured, human-managed data environments do not adequately account for the generative, probabilistic nature of AI-produced content. Specifically, dimensions such as hallucination susceptibility, bias amplification, and explainability are not captured by conventional quality models, highlighting the need to extend the theoretical landscape to accommodate the unique characteristics of GAIGS. Nevertheless, the recognition and confirmation of such factors lack consistent results in what has already been done, particularly in the implementation of the factors in the public sector [32]-[36]. This ambiguity and non-standardisation make it difficult to implement systematic evaluation or enhancement of the quality of the information produced by GAI in e-Government services. On the user side, information quality is one of the main predictors of satisfaction, behavioural intention, and future use of digital services [37], [38]. As soon as users believe that GAI-created content is truthful, understandable, and useful, their trust in working with the system rises considerably [33], [39]-[41]. On the other hand, low-quality or unclear content will build distrust and undermine the bond between citizens and digital government services.

Efforts to create GAI applications that are accountable and trusted by the citizens are not coherent without a firm base of clearly laid out information quality aspects. In addition, Chen and Wen [42], Afroogh et al. [43], and Zhu [44] pointed out that

the tendency to evaluate the quality of information in e-government supported by GAI is a complex task that reveals the urgency of determining the features appropriate in this changing environment. Quality information is also important in promoting initial satisfaction through delivering information of quality that users expect in terms of clarity, relevance, and reliability [45], [46]. Once the users are satisfied, they tend to have a positive behavioural intention towards using the service. Such intent is not only before actual usage but also enhances chances of making further usage in the future, as Prasetya et al. [47] elaborate. In GAI-supported platforms, Al-Haddad et al. [48] found that the behavioural intention is dependent not only on the technical accuracy of the information, but also on the capability of the user to interpret and contextualise the information. Such a systematic literature review is thus required to summarise the existing body of knowledge, find regularities, and suggest a consistent set of information quality indicators that could be used to effectively exploit GAI in e-Government environments. Taken together, the extant literature demonstrates a developing and divided knowledge on the quality of information in GAI-based e-Government services [49]. This gap is crucial to fill in order to advance the theoretical insights as well as the practical application of high-quality, reliable GAI-enabled public services.

III. Methodology

The study involved a systematic literature review methodology to determine, appraise, and summarise extant research regarding the determinants of the quality of information in GAIGS. This methodology was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines that have been developed by Moher et al. [50] and offer a standardised structure that guarantees methodological rigour, transparency, and replicability of a review. The process of selecting articles was recorded by the PRISMA 2009 flow diagram (see Fig. 1), which graphically shows the steps of identification, screening, eligibility, and inclusion.

The process was also divided into seven key phases, as suggested by Shaffril et al. [51], including the development of review protocols, the development of research questions, systematic searching tactics, quality evaluation, data extraction, data synthesis, and data demonstration. Other sources used in informing the methodology are Kitchenham and Charters [52], Petticrew and Roberts [53], and Hong et al. [54], which provide practical advice on the screening of articles and quality evaluation in non-health research studies.

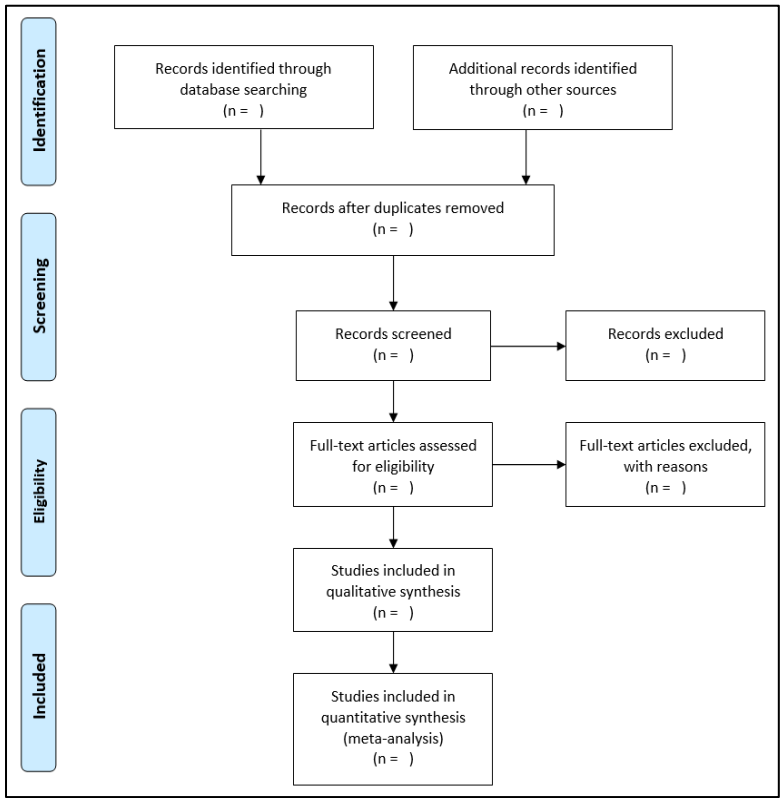


Fig. 1. PRISMA 2009 flow diagram for the systematic review process [50].

A. Review Protocol

As its main guideline during the systematic review process, the PRISMA reporting protocol [50] was also adopted. PRISMA is an established protocol for performing transparent and complete literature reviews in various academic sectors. Since the study is interdisciplinary in nature, additional information was also obtained in the form of software engineering and social science review guidelines to make the study domain relevant.

B. Formulation of Research Questions

This systematic review was developed using the PICo framework of research questions, which are more appropriate

in qualitative and exploratory reviews. PICo is an acronym of Population, Interest, and Context, a structured method of formulating clear, targeted, and answerable questions in evidence-based research [55]. Users and stakeholders of GAIGS are referred to as the Population in the study; the Interest is the key factor affecting the quality of information; the Context is the application of Generative AI technologies in e-government service environments. It is also important because PICo will make the objectives of the review clearer and relevant, which will ensure that the synthesised and extracted data answers the main questions that are being explored. Table I presents these research questions that were formulated using the PICo framework.

TABLE. I. RESEARCH QUESTIONS

No.	Question
RQ1	What are the key factors that influence information quality in Generative AI-enabled e-government Services (GAIGS)?
RQ2	How are these factors thematically categorised to reflect the current understanding of information quality in the context of eGovernment services?

C. Systematic Searching Strategies

In order to achieve a thorough and replicable evidence base, a systematic search strategy was used in the various academic databases. The approach was designed to identify peer-reviewed journal articles that address the intersection of information quality, Generative AI technologies, and e-government services. Following established systematic review protocols, the process involved clear stages of identification, screening, eligibility assessment, and quality appraisal, as detailed in the subsequent subsections.

1) Identification: The identification process involved a systematic search across four academic databases: Web of Science (WoS), IEEE Xplore, Scopus, and Wiley Online Library. The reason for selecting these databases was their ability to index a wide list of quality journals in the information systems, artificial intelligence, and technology in the public sector. To ensure comprehensive coverage, core keywords were first established, and related terms were further developed by consulting academic thesauri, domain-specific dictionaries, encyclopaedias, and relevant prior studies. This strategy enabled a thorough representation of key concepts related to information quality, Generative AI, and electronic government services. Boolean operators were applied to combine search terms strategically. The detailed search strings used for each database are presented in Table II. The initial search process yielded a total of 664 articles, comprising 114 from Web of Science, 122 from Scopus, 67 from IEEE Xplore, and 361 from Wiley Online Library.

TABLE. II. THE SEARCH STRINGS

Keywords	**Generative AI, e-Government Services, Information Quality, Factors, and Dimensions**
Database	**Search String**
Web of Science (WoS)	TS=("Information Quality" AND ("Generative AI" OR "Large Language Model*" OR "ChatGPT") AND ("e-Government" OR "electronic government" OR "digital government")) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article) AND PUBLICATION YEARS: (2021 OR 2022 OR 2023 OR 2024 OR 2025)
Scopus	TITLE-ABS-KEY (("Information Quality") AND ("Generative AI" OR "Large Language Model*" OR "ChatGPT") AND ("e-Government" OR "electronic government" OR "digital government")) AND PUBYEAR > 2020 AND PUBYEAR < 2026 AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (PUBSTAGE, "final")) AND (LIMIT-TO (SRCTYPE, "j"))
IEEE Xplore	"All Metadata":"Information Quality" AND ("Generative AI" OR "Large Language Model*" OR "ChatGPT") AND ("e-Government" OR "electronic government" OR "digital government") AND Publication Year: 2021–2025 AND Content Type: Journals AND Language: English
Wiley Online Library	("Information Quality") AND ("Generative AI" OR "Large Language Model*" OR "ChatGPT") AND ("e-Government" OR "electronic government" OR "digital government") AND Publication Date: 2021–2025 AND Publication Type: Journal Article AND Language: English

TABLE. III. THE SELECTION CRITERIA

Criterion	Inclusion	Exclusion
Language	English	Non-English
Timeline	2021 – 2025	< 2021
Literature type	Journal (Article)	Conference, Book, Review
Publication Stage	Final	In Press, Early Access

2) Screening: The screening process represents the initial stage of study selection in a systematic review. At this stage, all records retrieved from the database searches were first subjected to duplicate removal, resulting in the exclusion of 24 duplicate articles. The screening process applied predefined inclusion and exclusion criteria, specifically restricting the results to English-language journal articles published between 2021 and 2025, limited to final-stage publications. The detailed screening conditions are summarised in Table III. Following this screening phase, a total of 371 articles were deemed relevant and eligible for further assessment. This ensured that only studies meeting the specified screening criteria were carried forward.

3) Eligibility: After the screening process, the complete texts of the 371 articles that were retrieved were carefully analysed. This step was referred to as the eligibility assessment and sought to confirm that every article was in line with the inclusion criteria identified in the review protocol. The author and co-authors tested the objectives of the study, the context of the investigation, and the existence of the relevant factors, such as the quality of information, in GAIGS. Using this evaluation, 226 articles were eliminated since they did not qualify according to the inclusion criteria. In all, 121 articles were taken to the next phase to be evaluated in terms of quality.

D. Quality Assessment of the Selected Articles

The author and the co-authors independently evaluated the quality of the chosen articles, paying attention to the abstract, methodology, and main findings. In accordance with Petticrew and Roberts [53], each article was qualitatively evaluated and classified into one of three categories: high, moderate, or low

quality. Only articles rated as high or moderate were considered eligible for inclusion in the review. To ensure consistency and objectivity in assessment, this study adapted quality evaluation criteria based on the guidelines by Hong et al. [54]. The evaluation focused on five key questions:

For each criterion, the author and co-authors responded using one of three options: Yes, No, or Can't tell. Articles that fulfilled four or more of these criteria were categorised as high quality, while those that met exactly three were considered moderate quality. Articles meeting only one or two criteria were classified as low quality. To maintain the credibility of the findings, only studies rated as high or moderate quality were included in the synthesis process. Out of the 121 articles shortlisted for quality assessment, only those that met the required level of methodological soundness, specifically, those rated 4 or 5, advanced to the synthesis stage. A total of 33 articles satisfied these quality assessment criteria. These studies exhibited adequate rigour and relevance to the investigation of information quality in GAIGS. Articles that did not meet the quality threshold were excluded to ensure the reliability of the synthesis. The distribution of quality scores across all assessed articles is displayed in Fig. 2, which summarises the results of the quality appraisal conducted using a five-point scoring system based on the predefined criteria outlined in Table IV.

E. Data Extraction

Data extraction was conducted systematically using a structured data form. For each article, key details were recorded, including the author's name, year of publication, article title, the type of Generative AI technology used, and the category of e-government services. The extraction was conducted independently by the lead authors, with regular consultation and validation from the co-authors to ensure consistency and accuracy across all entries. For this study, a systematic integrative analysis served as the primary evaluation method, emphasising the synthesis of diverse research approaches, particularly those employing quantitative methods. The primary objective was to identify essential themes and subthemes, starting with the data collection phase, which presents an in-depth analysis of the 33 selected publications related to the study's focus. The subsequent analysis followed a thematic synthesis approach, with themes developed deductively based on pre-identified aspects of information quality and their intersection with Generative AI technologies in electronic government contexts. The research team cooperated in theme formulation, basing it on the evidence that was extracted.

TABLE. IV. QUALITY ASSESSMENT CRITERIA

No	Assessment Criteria
1	Are the main objectives of the article aligned with the investigation of information quality in Generative AI-enabled e-Government Services (GAIGS)?
2	Does the article clearly describe its research design and methodological approach?
3	Are the constructs or factors influencing information quality well-defined and contextually appropriate?
4	Is there sufficient justification for the selection and use of Generative AI technologies within e-Government settings?
5	Does the article provide practical insights or implications that contribute to the development or measurement of information quality in AI-enabled systems?

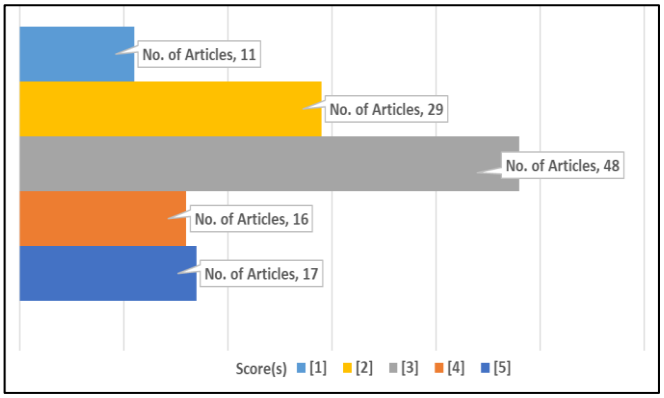


Fig. 2. Quality assessment results.

During the analysis, a reflective log was kept in order to capture different interpretations, difficulties, and emerging thoughts that are crucial in exploring data. All the differences or conceptual inconsistencies that were experienced throughout the process of developing the themes were addressed jointly with an aim of attaining coherence and reliability within the final thematic structure.

The expert review step also helped to improve the thematic framework. The suitability, the relevance, and the clarity of the identified subthemes were confirmed in this stage; therefore, the domain validity was established. Expert feedback was also included in order to improve the structure and make sure that it is consistent with the scope of the review.

F. Thematic Analysis

This study has established 22 important factors that determine the quality of information in GAIGS. This was based on a critical assessment of 33 chosen journal articles, and the results were analysed through thematic analysis on the basis of the six-step method suggested by Dawadi [56]. This method of analysis allowed the critical organisation of the factors into logical and sensible themes, providing a systematic meaning of how the quality of information is handled in the outputs of Generative AI in the context of delivering services to the people. The following is how it works:

In Step 1: Data Familiarisation, the process started with the repetition of the reading and reflecting on the 22 factors and how many times they appear in the 33 reviewed studies. Each factor was considered concerning its definition, role, and usage in the GAIGS context. This action made the research team form a general concept of the factors; hence, early conceptual overlaps and patterns could be identified, which would shape further analysis.

In Step 2: Generating Initial Codes, all the factors were given initial codes according to their conceptual contribution to the quality of information. As an illustration, credibility and reliability, factual correctness, and reference validity were paired with each other, whereas bias, transparency, and security were classified under ethical safeguards. These preliminary

codes were the role of each of these factors within the context of AI-generated content within the setting of public service.

In Step 3: Seeking Themes, the first codes were clustered into concept-similar categories, leading to the discovery of five main themes that reflect the key areas of concern with information quality in Generative AI applications. Trustworthiness and verifiability are the first theme and comprise reliability, factual accuracy, reference validity, traceability, verifiability, and authoritativeness. These considerations apply to the reliability and accuracy of GAI-created content and how users can challenge its authenticity, which is especially critical in the case of public services where misinformation may undermine institutional trust. The second theme, security and ethics, is composed of bias, transparency, security, compliance, and detectability. These considerations are an indication of the necessity of responsible system behaviour and ethical measures that safeguard users, foster fairness, and encourage accountability in the provision of AI services. The third theme, content quality and structure, includes such aspects as consistency, timeliness, completeness, interpretability, and granularity, which focus on the necessity of well-organised, clear, and timely content addressed to the informational needs of users. The fourth theme, user perception and value, entails relevance, accessibility, value-added, and uniqueness. These aspects demonstrate the user experiences and assessments of GAI-produced outputs with special reference to the utility, inclusiveness, and novelty of the information presented. Lastly, the fifth theme is adaptability and system behaviour, which encompasses hallucination and interoperability and touches on the ability of GAI systems to work on multiple platforms and act in response to various situations and handle risk arising out of content creation or inconsistency. Collectively, these themes give a complete picture of the landscape of factors that affect information quality in GAIGS.

In Step 4: Reviewing Themes, the research team re-examined every theme to achieve internal consistency and conceptual clarity. The individual factors were evaluated to ensure that they fit the targeted theme. As an example, hallucination was discussed in more than one category but eventually categorised under Adaptability and System Behaviour because of its direct relation to unpredictable GAI outputs. This action made respective themes unique and thematic boundaries clear.

In Step 5: Defining and Naming Themes, the five themes were narrowed down and were clearly named to indicate their focus on analysis. Trustworthiness and verifiability are used to describe the credibility and the reliability of GAI products. Security and ethics consist of measures that maintain fairness, transparency, and data protection. Content quality and structure measure the technical delivery and organisational clarity of the information. User perception and value are aspects of user-centred issues like usefulness and inclusiveness. Adaptability and system behaviour include the capability of the system to change with regard to user context and cross-platform functionality.

In the final step, Step 6: Producing the Report, the five themes identified were integrated into a coherent story that summarises the larger trends within the literature. Collectively, all these themes offer a complete picture of the various interrelated factors that constitute information quality in GAIGS. The thematic framework resulting from this process will serve as a starting point to further formulate the model and provide guidance to practitioners on how to design citizen-centred, effective, and trustworthy AI systems to serve the public sector.

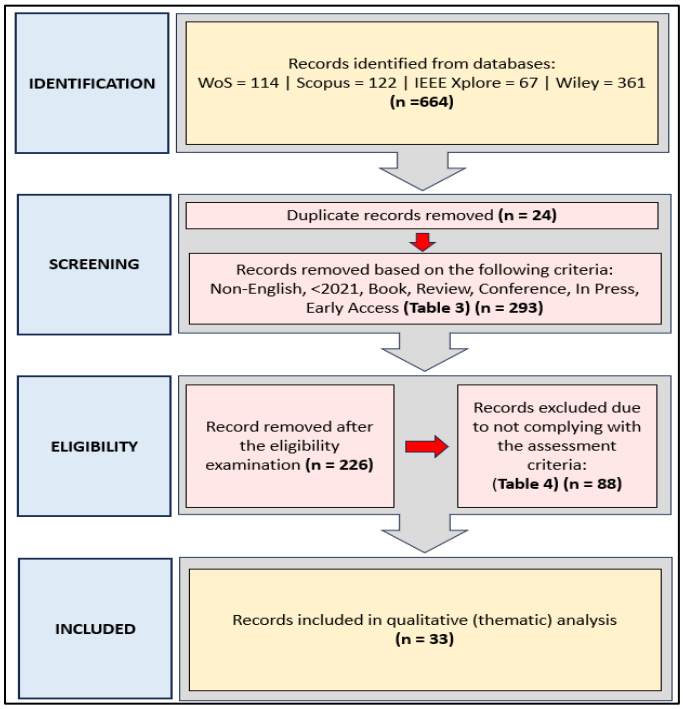


Fig. 3. PRISMA flow diagram applied for this study (adapted from [50]).

G. Data Demonstration

To maintain transparency and facilitate reproducibility, a PRISMA flow diagram was developed to illustrate the article identification, screening, eligibility, and inclusion stages, as depicted in Fig. 3. A supplementary table (Table VI) was also prepared to present the database results, screening outcomes, inclusion criteria, quality evaluation summary, and thematic classification of the 22 information quality factors. These visualisations support the interpretability and clarity of the systematic review, while providing a structured foundation for future empirical validation and model development.

IV. Results and Discussion

This section presents the findings derived from a systematic review and thematic analysis of 33 selected journal articles that focus on information quality in GAIGS. The results are synthesised through two key outputs. Table V provides an overview of the selected articles, detailing the types of Generative AI technologies examined and the corresponding context of e-government services. This table establishes the empirical foundation for the thematic synthesis.

Table VI consolidates the frequency and distribution of the 22 identified information quality factors, accompanied by references to the supporting articles. To provide quantitative grounding for the thematic findings, the frequency with which

each factor was cited across the 33 reviewed studies was recorded. Among the 22 identified factors, Timeliness emerged as the most frequently cited (n=10), followed by Reliability (n=8), Interpretability (n=8), Hallucination (n=8), Relevance (n=8), and Consistency (n=8). These high-frequency factors indicate areas of strongest consensus in the literature regarding what constitutes information quality in GAIGS. The authors further observed co-occurrence patterns in which Reliability, Factual Accuracy, and Transparency were repeatedly cited together, reinforcing their conceptual alignment within the Trustworthiness and Verifiability theme. Similarly, Hallucination and Bias were frequently co-cited, reflecting interrelated concerns about GAI output integrity. Together, these tables illustrate the diversity and recurrence of specific information quality concerns across different public sector domains, setting the stage for a deeper thematic exploration of how these factors cluster into broader conceptual themes.

TABLE. V. OVERVIEW OF SELECTED ARTICLES HIGHLIGHTING INFORMATION QUALITY FACTORS, GENERATIVE AI TYPES, AND E-GOVERNMENT SERVICE CONTEXT

References	Title	Type of GAI	e-G Services Context
[57]	Survey of Hallucination in Natural Language Generation	Natural Language Generation (NLG)	Healthcare, Education, Legal
[58]	Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References	Large Language Models (LLM)	Education
[59]	Hallucinations in ChatGPT: A Cautionary Tale for Biomedical Researchers	Large Language Models (LLM)	Healthcare
[60]	High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content	Large Language Models (LLM)	Healthcare
[61]	Comparing Scientific Abstracts Generated by ChatGPT to Real Abstracts with Detectors and Blinded Human Reviewers	Large Language Models (LLM)	Healthcare, Education
[62]	Solving Real-Time Information Updates and Mitigating Bias in Generative AI Models	General GAI	Public Information
[63]	Understanding and Addressing AI Hallucinations in Healthcare and Life Sciences	Large Language Models (LLM)	Healthcare
[64]	Bane and Boon of Hallucinations in the Context of Generative AI	General GAI	Public Information
[65]	Assessment of the Bias of Artificial Intelligence Generated Images and Large Language Models on Their Depiction of a Surgeon	Large Language Models (LLM)	Healthcare
[66]	An Analysis of the Accuracy and Bias of a Generative AI Model	General GAI	Public Information
[67]	From Automation to Innovation: Harnessing the Synergy of Generative AI and Retrieval-Augmented Generation for Citizen-Centric Governance	Retrieval-Augmented Generation (RAG) with Large Language Models (RAG-LLM)	Public Information
[68]	Improving citizen-government interactions with generative artificial intelligence: Novel human-computer interaction strategies for policy understanding through large language models	Human-Computer Interaction with Large Language Models (HCI–LLM)	Public Administration
[69]	Comparative Analysis of Generative AI Risks in the Public Sector	General GAI	Public Administration
[70]	Understanding the Interplay Between Trust, Reliability, and Human Factors in the Age of Generative AI	Human Factors and Large Language Models (LLM)	Public Administration
[71]	Tribal Knowledge Cocreation in Generative Artificial Intelligence Systems	Knowledge Co-Creation with Large Language Models (LLM)	Public Administration
[72]	Ethical Challenges and Solutions of Generative AI: An Interdisciplinary Perspective	AI Ethics and Natural Language Processing (NLP)	Public Administration
[73]	Elevating E-Government: Unleashing the Power of AI and IoT for Enhanced Public Services	General GAI	Public Administration
[74]	Beyond Algorithmic Disclosure for Generative AI	General GAI	Legal
[75]	Measuring Service Quality in Generative AI Environments: A Comprehensive GAISQUAL Framework	General GAI	Business, Information Science
[76]	Transparency for AI Systems: A Value-Based Approach	General GAI	Legal, Ethics
[77]	Generative Conversational AI Agent for Managerial Practices: The Role of IQ Dimensions, Novelty Seeking, and Ethical Concerns	Large Language Models (LLM)	Business, Management
[78]	Information Quality Dimensions in Generative Conversational AI for Financial Inclusion	General GAI	Financial Inclusion
[79]	Adoption challenges to artificial intelligence literacy in public healthcare: An evidence-based study in Saudi Arabia	Literacy Adoption and General GAI	Healthcare
[80]	Public opinion and the rise of digital minds: Perceived risk, trust, and regulation support	Governance, Public Trust and General GAI	Public Administration
[81]	Proficiency, Clarity, and Objectivity of Large Language Models Versus Specialists' Knowledge on COVID-19's Impacts in Pregnancy: Cross-Sectional Pilot Study	Large Language Models (LLM)	Healthcare
[82]	AI Can Be a Powerful Social Innovation for Public Health if Community Engagement Is at the Core	Natural Language Processing (NLP)	Public Health
[83]	Era of Generalist Conversational Artificial Intelligence to Support Public Health Communications	Large Language Models (LLM)	Public Health

References	Title	Type of GAI	e-G Services Context
[84]	Generative Artificial Intelligence (ChatGPT & Bard) in Public Administration Research: A Double-Edged Sword for Street-Level Bureaucracy Studies	Large Language Models (LLM)	Public Administration
[85]	Generative AI in Public Administration in Light of the Regulatory Awakening in the US and EU	General GAI	Public Administration
[86]	Unlocking the Power and Future Potential of Generative AI in Government Transformation	General GAI	Public Administration
[87]	Public Service with Generative AI: Exploring Features and Applications	General GAI	Public Administration
[88]	Evaluating public sector employee perceptions towards artificial intelligence and generative artificial intelligence integration	General GAI	Public Administration
[89]	E-Government 3.0: An AI Model to Use for Enhanced Local Democracies	General GAI	Public Administration, Local Governance

TABLE. VI. INFORMATION QUALITY FACTORS IDENTIFIED FROM SELECTED STUDIES

Factor	References
Reliability	[57], [58], [59], [70], [77], [83], [86], [89]
Factual Accuracy	[58]; [60], [62], [65], [68], [80], [82]
Interpretability	[57], [60], [63], [68], [69], [77], [83], [89]
Relevance	[58], [59], [61], [65], [66], [70], [74], [82]
Timeliness	[58], [59], [61], [64], [69], [72], [74], [79], [80], [82]
Consistency	[57], [60], [61], [65], [68], [71], [72], [77]
Transparency	[60], [61], [65], [70], [71], [75], [83]
Hallucination	[57], [59], [60], [75], [80], [81], [83], [87]
Bias	[58], [62], [66], [67], [81], [84]
Granularity	[65], [72], [73], [81], [85]
Accessibility	[59], [67], [69], [83], [85]
Security	[59], [60], [76], [81], [82]
Completeness	[73], [78], [84]
Compliance	[61], [67], [75], [79], [86]
Value-Added	[59], [80], [83], [88]
Detectability	[57]; [62], [65], [72], [79]
Traceability	[60], [62], [65], [85]
Authoritativeness	[58]; [64], [66], [82]
Uniqueness	[58], [65], [71]
Interoperability	[62], [80]
Verifiability	[60], [66], [67]
Reference Validity	[64], [84]

The themes identified in this review, as depicted in Fig. 4, are strongly supported by recurring patterns across the selected studies on Generative AI in e-government services. The need for trust, factual accuracy, and source validation is repeatedly emphasised in GAI applications deployed in domains such as healthcare, education, and public administration [57], [58], [60], [62]. These studies reinforce the critical role of factors under the theme of trustworthiness and verifiability in ensuring the credibility of GAI-generated content. The perceived reliability of the system decreases when trust is violated, and this factor can adversely affect the user's desire to embrace or trust GAI to deliver critical services to society. Likewise, the theme of security and ethics is supported in findings that investigate the risks of bias, lack of transparency, and non-adherence in outputs produced by GAI [65], [84], [85]. Such ethical issues are especially of concern in the domain of e-government, where accountability and equity are of paramount importance in terms of system acceptance by the populace.

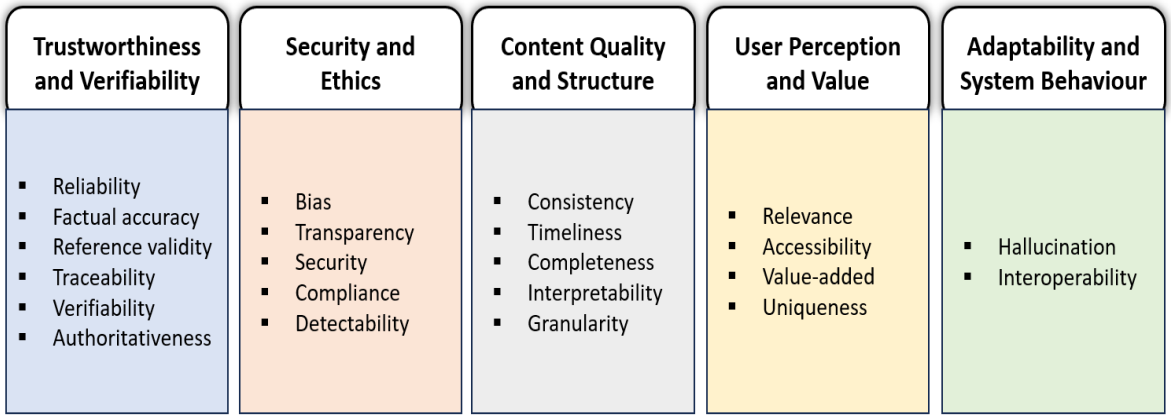


Fig. 4. Themes of factors influencing information quality in generative AI-enabled e-government services (GAIGS).

Moreover, the concept of content quality and structure is complemented by the findings that emphasise the necessity of consistency, timeliness, and interpretability to make the information actionable and the interactions in the service context smooth [68], [69]. The GAI systems that provide clear and well-structured information regularly are more likely to create positive perceptions that may contribute to the increase in the intention to use and further use such services. A similar user perception and value theme is supported by the literature that refers to the role of accessibility, relevance, and usefulness in satisfaction with e-government platforms [12], [59], [81]. These are key elements in creating confidence in GAI-assisted interactions and maintaining interest on the part of users. Furthermore, the topic of adaptability and system behaviour is confirmed by articles on hallucination, system integration, and context-sensitive output requirements [83]. Such features affect not only the current usability but also the long-term confidence in the work of the system. Collectively, the results indicate that the quality of GAI-generated information plays an important role in influencing the initial user acceptance as well as the intent to use the GAI-enabled eGovernment services further [90]. Based on the synthesised evidence, the authors propose the following system design guidelines for GAIGS practitioners. First, for Trustworthiness and Verifiability, platforms should incorporate source-attribution layers and explainability mechanisms that enable users to trace and verify the provenance of AI-generated content. Second, for Security and Ethics, developers are advised to embed bias-detection pipelines and conduct periodic algorithmic audits, supported by compliance monitoring aligned to national e-government standards. Third, for Content Quality and Structure, knowledge-grounding mechanisms should be adopted to reduce hallucination and improve the timeliness and consistency of outputs. Fourth, for User Perception and Value, accessibility-by-design principles, including multilingual support and plain-language generation, should be applied to serve diverse citizen populations. Fifth, for Adaptability and System Behaviour, systems should be designed with context-sensitive response tuning and cross-platform interoperability standards to ensure consistent quality across different service delivery channels.

V. Conclusion

The review has synthesised evidence of thirty-three identified articles published in 2021-2025 to determine twenty-two main factors contributing to information quality in GAI-enabled eGovernment services (GAIGS). Through a structured thematic analysis, these factors were grouped into five major themes, namely trustworthiness and verifiability, security and ethics, content quality and structure, user perception and value, and adaptability and system behaviour. The themes capture the varying methods in the conceptualisation of the information quality of the AI-generated outputs of public services in the literature. It is shown in the findings that credible, ethical, convenient, and contextually responsive information is critical to the improvement of user satisfaction and to the desire and persistence in the use of GAI-assisted platforms.

Several limitations of this review should be acknowledged. The corpus is restricted to peer-reviewed journal articles, which may exclude relevant insights available in grey literature and government reports. The thematic classification process, while guided by an established six-step method, involves a degree of interpretive judgement. Furthermore, given the rapid pace of GAI development, emerging capabilities may introduce quality dimensions not yet captured in the existing literature. Notwithstanding these limitations, the authors identify three priority directions to guide future inquiry. First, the 22 factors identified in this review should be subjected to empirical validation through the development and psychometric testing of measurement instruments in real GAIGS deployment contexts. Second, domain-specific information quality models should be constructed for distinct e-government service areas, recognising that quality priorities may vary considerably across healthcare, legal, taxation, and social welfare services. Third, future research should examine user-side moderating variables, including digital literacy, trust disposition, and prior experience with AI systems, to better understand how these factors mediate the relationship between information quality and citizen satisfaction, adoption, and continued engagement. Despite these limitations, the authors are confident that the findings of this review offer a timely, structured, and practical foundation for advancing both the theoretical understanding and the responsible implementation of high-quality, trustworthy, and citizen-centred GAIGS.

Acknowledgment

This research was funded by the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia.

References

[1] J. Longo, "The transformative potential of artificial intelligence for public sector reform," Canadian Public Administration, Dec. 2024. [Online]. Available: https://doi.org/10.1111/capa.12587.

[2] M. Crispin-Ortuzar and N. E. Weckman, "Personalised, connected health enabled by AI and home-based diagnostics," Trends in Biotechnology, vol. 41, no. 7, pp. 982–983, 2023. [Online]. Available: https://doi.org/10.1016/j.tibtech.2023.03.013.

[3] Y. Wang, Y. Pan, M. Yan, Z. Su, and T. H. Luan, "A survey on ChatGPT: AI-generated contents, challenges, and solutions," IEEE Open Journal of the Computer Society, vol. 4, pp. 280–302, 2023. [Online]. Available: https://doi.org/10.1109/OJCS.2023.3300321.

[4] D. Dalalah, "The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT," Int. J. Manag. Educ., vol. 21, no. 2, Art. no. 100822, 2023. [Online]. Available: https://doi.org/10.1016/j.ijme.2023.100822.

[5] P. Hacker, A. K. Engel, and M. Mauer, "Regulating ChatGPT and other Large Generative AI Models," in Proc. 2023 ACM Conf. Fairness, Accountability, and Transparency, pp. 1112–1123, Feb. 2023. [Online]. Available: https://doi.org/10.1145/3593013.3594067.

[6] C. van Noordt and G. Misuraca, "Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the European Union," Government Information Quarterly, vol. 39, no. 3, Art. no. 101714, 2022. [Online]. Available: https://doi.org/10.1016/j.giq.2022.101714.

[7] P. Robles and D. J. Mallinson, "Catching up with AI: Pushing toward a cohesive governance framework," Politics & Policy, vol. 51, no. 3, pp. 355–372, 2023. [Online]. Available: https://doi.org/10.1111/polp.12529.

[8] P. S. Varsha, A. Chakraborty, and A. K. Kar, "How to undertake an impactful literature review: Understanding review approaches and guidelines for high-impact systematic literature reviews," J. Inf. Sci., vol. 50, no. 3, pp. 1–24, 2024. [Online]. Available: https://doi.org/10.1007/978-3-031-22950-3_9.

[9] A. P. Manoharan, J. Melitski, and M. Holzer, "Digital governance: An assessment of performance and best practices," Public Organization

Review, vol. 23, pp. 265–283, 2023. [Online]. Available: https://doi.org/10.1007/s11115-021-00584-8.

[10] D. Wa and Z. Zhang, "Research on e-government adoption in environmental governance from the perspective of public participation: Empirical analysis based on 31 provinces in China," Pol. J. Environ. Stud., vol. 32, no. 2, pp. 1383–1391, 2023. [Online]. Available: https://doi.org/10.15244/pjoes/153076.

[11] D. E. A. Danu, S. Supardi, S. T. J. P. Sutanto, and R. S. Riko, "E-services: Implementation of digital-based public services in the 4.0 era," Athena: J. Social, Culture and Society, vol. 1, no. 3, pp. 87–92, 2023. [Online]. Available: https://doi.org/10.58905/athena.v1i3.40.

[12] I. K. Mensah and D. S. Mwakapesa, "The impact of e-government information quality (EGIQ) dimensions on the adoption of electronic government services," Inf. Dev., vol. 41, no. 2, pp. 265–284, 2023. [Online]. Available: https://doi.org/10.1177/02666669231155164.

[13] U. Hidayati et al., "The effect of system reliability, information sharing and service quality on e-learning net benefit in public sector organisations," Int. J. Data Netw. Sci., vol. 7, no. 3, pp. 1397–1404, 2023. [Online]. Available: https://doi.org/10.5267/j.ijdns.2023.3.024.

[14] M. Binz and E. Schulz, "Using cognitive psychology to understand GPT-3," Proc. Natl. Acad. Sci. U.S.A., vol. 120, no. 6, 2023. [Online]. Available: https://doi.org/10.1073/pnas.2218523120.

[15] D. M. Markowitz, J. T. Hancock, and J. N. Bailenson, "Linguistic markers of inherently false AI communication and intentionally false human communication: Evidence from hotel reviews," J. Lang. Soc. Psychol., vol. 43, no. 1, pp. 63–82, 2023. [Online]. Available: https://doi.org/10.1177/0261927x231200201.

[16] M. Jakesch, J. T. Hancock, and M. Naaman, "Human heuristics for AI-generated language are flawed," Proc. Natl. Acad. Sci. U.S.A., vol. 120, no. 11, 2023. [Online]. Available: https://doi.org/10.1073/pnas.2208839120.

[17] V. J. Straub, D. Morgan, J. Bright, and H. Margetts, "Artificial intelligence in government: Concepts, standards, and a unified framework," Gov. Inf. Q., vol. 40, no. 4, Art. no. 101881, 2023. [Online]. Available: https://doi.org/10.1016/j.giq.2023.101881.

[18] A. J. Desmal, M. K. B. Othman, S. Hamid, and A. H. S. Zolait, "Exploring the information quality of mobile government services: A literature review," PeerJ Comput. Sci., vol. 8, Art. no. e1028, 2022. [Online]. Available: https://doi.org/10.7717/peerj-cs.1028.

[19] M. Moreno Ibarra, M. Saldaña Pérez, S. Pérez Rodríguez, and E. Juárez Carbajal, "Generative AI (GenAI) and smart cities: Efficiency, cohesion, and sustainability," in Handbook of Smart City Innovations, Routledge, 2024, ch. 11. [Online]. Available: https://doi.org/10.1201/9781003415930-11.

[20] F. Leiser, S. Eckhardt, V. Leuthe, M. Knaeble, A. Mädche, G. Schwabe, and A. Sunyaev, "HILL: A hallucination identifier for large language models," in CHI '24: Proc. 2024 CHI Conf. Human Factors Comput. Syst., Art. no. 482, pp. 1–13, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642428.

[21] Y. Zhang and R. Gosline, "Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation," Judgm. Decis. Mak., vol. 18, 2023. [Online]. Available: https://doi.org/10.1017/jdm.2023.37.

[22] M. Temsah et al., "ChatGPT and the future of digital health: A study on healthcare workers' perceptions and expectations," Healthcare, vol. 11, no. 13, Art. no. 1812, 2023. [Online]. Available: https://doi.org/10.3390/healthcare11131812.

[23] Y. Sun, D. Sheng, Z. Zhou, and Y. Wu, "AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content," Humanit. Soc. Sci. Commun., vol. 11, Art. no. 1278, 2024. [Online]. Available: https://doi.org/10.1057/s41599-024-03811-x.

[24] S. Banerjee et al., "A clinical insight on new discovered molecules and repurposed drugs for the treatment of COVID-19," Vaccines, vol. 11, no. 2, Art. no. 332, 2023. [Online]. Available: https://doi.org/10.3390/vaccines11020332.

[25] H. Torkamaan et al., "Challenges and future directions for integration of large language models into sociotechnical systems," Behav. Inf. Technol., pp. 1–20, 2024. [Online]. Available: https://doi.org/10.1080/0144929X.2024.2431068.

[26] Q. Huang, C. Lv, L. Lu, and S. Tu, "Evaluating the quality of AI-generated digital educational resources for university teaching and learning," Systems, vol. 13, no. 3, Art. no. 174, 2025. [Online]. Available: https://doi.org/10.3390/systems13030174.

[27] I. Y. Zhao et al., "Ethics, integrity, and retributions of digital detection surveillance systems for infectious diseases: Systematic literature review," J. Med. Internet Res., vol. 23, no. 10, Art. no. e32328, 2021. [Online]. Available: https://doi.org/10.2196/32328.

[28] M. Hutahaean, I. J. Eunike, and A. D. K. Silalahi, "Do social media, good governance, and public trust increase citizens' e-government participation? Dual approach of PLS-SEM and fsQCA," Hum. Behav. Emerg. Technol., 2023, pp. 1–19. [Online]. Available: https://doi.org/10.1155/2023/9988602.

[29] M. Esposito and T. Tse, "Mitigating the risks of generative AI in government through algorithmic governance," in Proc. ACM Digit. Gov. Conf., pp. 67–81, 2024. [Online]. Available: https://doi.org/10.1145/3657054.3657124.

[30] A. M. G. Zuiderwijk, Y.-C. Chen, and F. Salem, "Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda," Gov. Inf. Q., vol. 38, no. 3, Art. no. 101577, 2021. [Online]. Available: https://doi.org/10.1016/j.giq.2021.101577.

[31] A. Lovari and F. De Rosa, "Exploring the challenges of generative AI on public sector communication in Europe," Media Commun., vol. 13, Art. no. 9644, 2025. [Online]. Available: https://doi.org/10.17645/mac.9644.

[32] O. D. Lee, R. Ayyagari, F. Nasirian, and M. Ahmadian, "Role of interaction quality and trust in use of AI-based voice-assistant systems," J. Syst. Inf. Technol., vol. 23, no. 2, pp. 154–170, 2021. [Online]. Available: https://doi.org/10.1108/jsit-07-2020-0132.

[33] M. Sallam, M. Barakat, and M. Sallam, "Pilot testing of a tool to standardise the assessment of the quality of health information generated by artificial intelligence-based models," Cureus, vol. 15, no. 11, 2023. [Online]. Available: https://doi.org/10.7759/cureus.49373.

[34] A. Doyal, D. Sender, M. Nanda, and R. Serrano, "Chat GPT and artificial intelligence in medical writing: Concerns and ethical considerations," Cureus, vol. 15, no. 8, 2023. [Online]. Available: https://doi.org/10.7759/cureus.43292.

[35] N. Aoki, T. Tatsumi, G. Naruse, and K. Maeda, "Explainable AI for government: Does the type of explanation matter to the accuracy, fairness, and trustworthiness of an algorithmic decision as perceived by those who are affected?" Gov. Inf. Q., vol. 41, no. 4, Art. no. 101965, 2024. [Online]. Available: https://doi.org/10.1016/j.giq.2024.101965.

[36] A. F. Vatamanu and M. Tofan, "Integrating Artificial Intelligence into Public Administration: Challenges and Vulnerabilities," Adm. Sci., vol. 15, no. 4, Art. no. 149, 2025. [Online]. Available: https://doi.org/10.3390/admsci15040149.

[37] A. Kulal, H. Ur Rahiman, H. Suvarna, N. Abhishek, and S. Dinesh, "Enhancing public service delivery efficiency: Exploring the impact of AI," J. Open Innov.: Technol., Mark., Complex., vol. 10, no. 3, Art. no. 100329, 2024. [Online]. Available: https://doi.org/10.1016/j.joitmc.2024.100329.

[38] A. D. K. Silalahi, "Can generative artificial intelligence drive sustainable behavior? A consumer adoption model for AI driven sustainability recommendations," Technol. Soc., Art. no. 102995, 2025. [Online]. Available: https://doi.org/10.1016/j.techsoc.2025.102995.

[39] Y. K. Ghanem et al., "Dr. Google to Dr. ChatGPT: Assessing the content and quality of artificial intelligence-generated medical information on appendicitis," Surg. Endosc., vol. 38, no. 5, pp. 2887–2893, 2024. [Online]. Available: https://doi.org/10.1007/s00464-024-10739-5.

[40] T. J. Lee et al., "Evaluating ChatGPT responses on atrial fibrillation for patient education," Cureus, vol. 16, no. 6, 2024. [Online]. Available: https://doi.org/10.7759/cureus.61680.

[41] F. J. Selten and B. Klievink, "Organising public sector AI adoption: Navigating between separation and integration," Gov. Inf. Q., vol. 41, no. 1, Art. no. 101885, 2024. [Online]. Available: https://doi.org/10.1016/j.giq.2023.101885.

[42] Y.-N. K. Chen and C.-H. R. Wen, "Impacts of attitudes toward government and corporations on public trust in artificial intelligence," Commun. Stud., vol. 72, no. 1, pp. 115–131, 2021. [Online]. Available: https://doi.org/10.1080/10510974.2020.1807380.

[43] S. Afroogh, A. Akbari, E. Malone, M. Kargar, and H. Alambeigi, "Trust in AI: Progress, challenges, and future directions," Humanit. Soc. Sci. Commun., vol. 11, Art. no. 1568, 2024. [Online]. Available: https://doi.org/10.1057/s41599-024-04044-8.

[44] J. Zhu, "Generative AI for content discovery in academic publishing: A content provider's perspective," Inf. Serv. Use, vol. 45, no. 1–2, pp. 48–55, 2025. [Online]. Available: https://doi.org/10.1177/18758789251336075.

[45] A. Hooda, P. Gupta, A. Jeyaraj, and Y. K. Dwivedi, "Clarifying the role of e-government trust in e-government success models: A meta-analytic structural equation modeling approach," Australas. J. Inf. Syst., vol. 27, 2023. [Online]. Available: https://doi.org/10.3127/ajis.v27i0.4079.

[46] H. Nam, T. Nam, and S. Kim, "Identifying the determinants of platform-based e-government service use," J. Glob. Inf. Manag., vol. 32, no. 1, pp. 1–21, 2024. [Online]. Available: https://doi.org/10.4018/jgim.336554.

[47] D. A. Prasetya, A. P. Sari, P. A. Riyantoko, and T. M. Fahrudin, "The effect of information quality and service quality on user satisfaction of the Government of Kabupaten Malang," TIERS Inf. Technol. J., vol. 4, no. 1, pp. 32–42, 2023. [Online]. Available: https://doi.org/10.38043/tiers.v4i1.4328.

[48] S. Al-Haddad, A. A. Sharabati, M. H. A. Khasawneh, S. A. Mazahreh, and Y. T. Kawar, "Behavioral acceptance of electronic government in Jordan," Int. J. Electron. Gov. Res., vol. 19, no. 1, pp. 1–26, 2023. [Online]. Available: https://doi.org/10.4018/ijegr.321459.

[49] T. T. U. Nguyen, P. V. Nguyen, H. T. N. Huynh, D. Vrontis, and Z. U. Ahmed, "Identification of the determinants of public trust in e-government services and participation in social media based on good governance theory and the technology acceptance model," J. Asia Bus. Stud., vol. 18, no. 1, pp. 44–61, 2024. [Online]. Available: https://doi.org/10.1108/JABS-04-2023-0160.

[50] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and The PRISMA Group, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," PLoS Med., vol. 6, no. 7, Art. no. e1000097, 2009. [Online]. Available: https://doi.org/10.1371/journal.pmed.1000097.

[51] H. A. M. Shaffril, S. F. Samsuddin, and A. Abu Samah, "The ABC of systematic literature review: A basic methodological guidance for beginners," Qual. Quant., vol. 55, pp. 1319–1346, 2020. [Online]. Available: https://doi.org/10.1007/s11135-020-01059-6.

[52] B. Kitchenham and S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering (EBSE 2007-001), Keele University and University of Durham, 2007.

[53] M. Petticrew and H. Roberts, Systematic Reviews in the Social Sciences: A Practical Guide. Oxford, UK: Blackwell Publishing, 2006.

[54] Q. N. Hong, P. Pluye, M. Bujold, and M. Wassef, "Convergent and sequential synthesis designs: Implications for conducting and reporting systematic reviews of qualitative and quantitative evidence," Syst. Rev., vol. 6, no. 1, Art. no. 61, 2018. [Online]. Available: https://doi.org/10.1186/s13643-017-0454-2.

[55] Joanna Briggs Institute, Joanna Briggs Institute Reviewer's Manual: Methodology for JBI Qualitative Systematic Reviews, The Joanna Briggs Institute, 2017. [Online]. Available: https://jbi-global-wiki.refined.site/space/MANUAL.

[56] S. Dawadi, "Thematic analysis approach: A step-by-step guide for ELT research practitioners," J. NELTA, vol. 25, no. 1–2, pp. 62–71, 2020.

[57] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Comput. Surv., vol. 55, no. 12, Art. no. 248, 2023. [Online]. Available: https://doi.org/10.1145/3571730.

[58] S. A. Athaluri, S. V. Manthena, M. K. Kesapragada, V. Yarlagadda, T. Dave, and A. Sinha, "Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references," Cureus, vol. 15, no. 4, Art. no. e37432, 2023. [Online]. Available: https://doi.org/10.7759/cureus.37432.

[59] J. Goddard, "Hallucinations in ChatGPT: A cautionary tale for biomedical researchers," Am. J. Med., vol. 136, no. 11, pp. 1059–1060, 2023. [Online]. Available: https://doi.org/10.1016/j.amjmed.2023.06.012.

[60] M. Bhattacharyya, V. M. Miller, D. Bhattacharyya, and L. E. Miller, "High rates of fabricated and inaccurate references in ChatGPT-generated medical content," Cureus, vol. 15, no. 5, Art. no. e39238, 2023. [Online]. Available: https://doi.org/10.7759/cureus.39238.

[61] C. A. Gao et al., "Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers," npj Digit. Med., vol. 6, no. 1, Art. no. 75, 2023. [Online]. Available: https://doi.org/10.1038/s41746-023-00819-6.

[62] A. Shah and V. Vishwakarma, "Solving real-time information updates and mitigating bias in generative AI models," Int. J. Innov. Sci. Res. Technol., vol. 9, no. 7, 2024. [Online]. Available: https://doi.org/10.38124/ijisrt/IJISRT24JUL1248.

[63] A. Gadiko, "Understanding and addressing AI hallucinations in healthcare and life sciences," Int. J. Health Sci., vol. 7, no. 3, pp. 1–11, 2024. [Online]. Available: https://doi.org/10.47941/ijhs.1862.

[64] S. M. N. Sakib, "Bane and boon of hallucinations in the context of generative AI," in Cases on AI Ethics in Business, IGI Global, 2024, pp. 276–299. [Online]. Available: https://doi.org/10.4018/979-8-3693-2643-5.ch016.

[65] J. Cevik et al., "Assessment of the bias of artificial intelligence generated images and large language models on their depiction of a surgeon," ANZ J. Surg., vol. 94, no. 3, pp. 287–294, 2024. [Online]. Available: https://doi.org/10.1111/ans.18792.

[66] C. Hunkele and V. Smith, "An analysis of the accuracy and bias of a generative AI model," Int. J. Adv. Res., vol. 12, no. 4, pp. 989–1019, 2024. [Online]. Available: https://doi.org/10.21474/ijar01/18640.

[67] S. Morande, G. Sogani, and S. Shah, "From automation to innovation: Harnessing the synergy of generative AI and retrieval-augmented generation for citizen-centric governance," in Proc. ITU Kaleidoscope 2024 Conf., pp. 1–8. IEEE, 2024. [Online]. Available: https://doi.org/10.23919/ituk62727.2024.10772932.

[68] L. Yun, S. Yun, and H. Xue, "Improving citizen-government interactions with generative artificial intelligence: Novel human-computer interaction strategies for policy understanding through large language models," PLOS ONE, vol. 19, no. 12, Art. no. e0311410, 2024. [Online]. Available: https://doi.org/10.1371/journal.pone.0311410.

[69] M. A. Beltran, M. I. Ruiz Mondragon, and S. H. Han, "Comparative analysis of generative AI risks in the public sector," in Proc. 25th Annu. Int. Conf. Digit. Gov. Res. (dg.o 2024), pp. 610–617. ACM, 2024. [Online]. Available: https://doi.org/10.1145/3657054.3657125.

[70] S. Thorne, "Understanding the interplay between trust, reliability, and human factors in the age of generative AI," Int. J. Simul.: Syst., Sci. Technol., vol. 25, no. 1, Art. no. 10, 2024. [Online]. Available: https://doi.org/10.5013/IJSSST.a.25.01.10.

[71] Y. Wang and J. Smith, "Tribal knowledge cocreation in generative artificial intelligence systems," in Proc. 25th Annu. Int. Conf. Digit. Gov. Res., pp. 637–644, ACM, 2024. [Online]. Available: https://doi.org/10.1145/3657054.3657129.

[72] M. Al-Kfairy, D. Mustafa, N. Kshetri, M. Insiew, and O. Alfandi, "Ethical challenges and solutions of generative AI: An interdisciplinary perspective," Informatics, vol. 11, no. 3, Art. no. 58, 2024. [Online]. Available: https://doi.org/10.3390/informatics11030058.

[73] A. M. Al-Ansi, A. Garad, M. Jaboob, and A. Al-Ansi, "Elevating e-government: Unleashing the power of AI and IoT for enhanced public services," Heliyon, vol. 10, no. 23, Art. no. e40591, 2024. [Online]. Available: https://doi.org/10.1016/j.heliyon.2024.e40591.

[74] C. Yoo, "Beyond algorithmic disclosure for generative AI," Sci. Technol. Law Rev., vol. 25, no. 2, 2024. [Online]. Available: https://doi.org/10.52214/stlr.v25i2.12766.

[75] J. G. Sankar, "Measuring service quality in generative AI environments: A comprehensive GAISQUAL framework," in Design and Validation of Research Tools and Methodologies, IGI Global, 2024, pp. 311–340. [Online]. Available: https://doi.org/10.4018/979-8-3693-1135-6.ch015.

[76] S. Buijsman, "Transparency for AI systems: A value-based approach," Ethics Inf. Technol., vol. 26, Art. no. 34, 2024. [Online]. Available: https://doi.org/10.1007/s10676-024-09770-w.

[77] A. M. Baabdullah, "Generative conversational AI agent for managerial practices: The role of IQ dimensions, novelty seeking, and ethical concerns," Technol. Forecast. Soc. Change, vol. 198, Art. no. 122951,

2024. [Online]. Available: https://doi.org/10.1016/j.techfore.2023.122951.

[78] K. Patil, D. Pramod, and M. S. Kulkarni, "Information quality dimensions in generative conversational AI for financial inclusion," in Proc. 6th Int. Conf. Comput. Intell. Commun. Technol. (CCICT), IEEE, 2024. [Online]. Available: https://doi.org/10.1109/CCICT62777.2024.00021.

[79] R. Kumar, A. Singh, A. Kassar, M. Humaida, S. Joshi, and M. Sharma, "Adoption challenges to artificial intelligence literacy in public healthcare: An evidence-based study in Saudi Arabia," Front. Public Health, vol. 13, 2025. [Online]. Available: https://doi.org/10.3389/fpubh.2025.1558772.

[80] J. Bullock, J. Pauketat, H. Huang, Y. Wang, and J. Anthis, "Public opinion and the rise of digital minds: Perceived risk, trust, and regulation support," Public Perform. Manag. Rev., pp. 1–32, 2025. [Online]. Available: https://doi.org/10.1080/15309576.2025.2495094.

[81] N. L. Bragazzi et al., "Proficiency, clarity, and objectivity of large language models versus specialists' knowledge on COVID-19's impacts in pregnancy: Cross-sectional pilot study," JMIR Form. Res., vol. 9, 2025. [Online]. Available: https://doi.org/10.2196/56126.

[82] A. N. Bazzano, A. Mantsios, N. Mattei, M. R. Kosorok, and A. Culotta, "AI can be a powerful social innovation for public health if community engagement is at the core," J. Med. Internet Res., vol. 27, 2025. [Online]. Available: https://doi.org/10.2196/68198.

[83] E. Sezgin and A. Kocaballi, "Era of generalist conversational artificial intelligence to support public health communications," J. Med. Internet Res., vol. 27, 2025. [Online]. Available: https://doi.org/10.2196/69007.

[84] M. Salah, F. Abdelfattah, and H. Al Halbusi, "Generative Artificial Intelligence (ChatGPT & Bard) in public administration research: A double-edged sword for street-level bureaucracy studies," Int. J. Public Admin., pp. 1–7, 2023. [Online]. Available: https://doi.org/10.1080/01900692.2023.2274801.

[85] S. Weerts, "Generative AI in public administration in light of the regulatory awakening in the US and EU," Cambridge Forum on AI: Law and Governance, vol. 1, Art. no. e3, 2025. [Online]. Available: https://doi.org/10.1017/cfl.2024.10.

[86] J. K. Pandey, "Unlocking the power and future potential of generative AI in government transformation," Transform. Gov.: People, Process Policy, 2024. [Online]. Available: https://doi.org/10.1108/TG-01-2024-0006.

[87] H. Aryfiyanto and A. Alamsyah, "Public service with generative AI: Exploring features and applications," in Proc. 7th Int. Conf. Comput. Informatics Eng. (IC2IE), pp. 1–7, IEEE, 2024. [Online]. Available: https://doi.org/10.1109/IC2IE63342.2024.10747963.

[88] L. Giraldi, L. Rossi, and E. Rudawska, "Evaluating public sector employee perceptions towards artificial intelligence and generative artificial intelligence integration," J. Inf. Sci., 2024. [Online]. Available: https://doi.org/10.1177/01655515241293775.

[89] C. Vrabie, "E-Government 3.0: An AI model to use for enhanced local democracies," Sustainability, vol. 15, no. 12, Art. no. 9572, 2023. [Online]. Available: https://doi.org/10.3390/su15129572.

[90] S. Bohara, G. Gupta, and V. Kumar, "Integrating artificial intelligence and consumer behavior: Insights and directions from a hybrid systematic literature review," J. Market. Commun., pp. 1–30, 2025. [Online]. Available: https://doi.org/10.1080/13527266.2025.2508181.

Balancing Accuracy Robustness and Explainability in E-Commerce Recommender Systems

Mansor Alohali
Applied College, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

***Abstract*—Recommender systems are essential to digital marketplaces, shaping how users discover products and engage with platforms. While AI has significantly improved accuracy, critical concerns about robustness and explainability remain. This study introduces and empirically validates the "Recommender's Trilemma"—an inherent trade-off between accuracy, robustness, and explainability. Through comparative analysis of NeuMF, SVD, and TF-IDF on the Amazon Electronics dataset, we uncover a dual failure cascade: adversarial attacks not only degrade recommendation quality but also destabilize the explanations meant to foster user trust. While NeuMF achieves high accuracy, it is susceptible to data poisoning that undermines its decision logic; in contrast, the transparent TF-IDF model offers interpretability but suffers from low predictive power and brittle explanations. These findings expose a structural vulnerability in recommender system design and provide a diagnostic framework for auditing deployed systems. We call for a new development paradigm where robustness and explainability are treated as co-primary objectives alongside accuracy—enabling trustworthy, resilient, and ethically aligned AI in digital commerce.**

***Keywords*—Recommender systems; E-commerce; explainable AI; adversarial robustness; personalization; digital platforms**

I. Introduction

Recommender systems are foundational to the digital economy, shaping user experiences across e-commerce, streaming, news, and social media platforms [1]. These systems leverage vast amounts of user data such as ratings, clicks, browsing behavior, and reviews [2]. As a result, recommender systems deliver personalized content that drives engagement, increases conversion rates, and enhances user retention [3]. As recommender systems become central to digital commerce ecosystems, concerns about their fairness, transparency, and resilience are increasingly relevant for both platform managers and policy makers. [4].

Despite rising awareness of ethical and operational concerns, the predominant evaluation paradigm for recommender systems remains narrowly focused on predictive accuracy [5]. Metrics such as Hit Rate (HR@k) [6], [7], Normalized Discounted Cumulative Gain (NDCG) [8], and Mean Reciprocal Rank (MRR) [9] are routinely used to assess how effectively a model ranks relevant items [10]. However, these metrics offer only a limited perspective: they say little about how resilient a model is to adversarial manipulation or whether the reasons behind its recommendations can be trusted by end-users, regulators, or developers [11]. This limitation is especially critical in e-commerce, where recommender systems not only drive sales but also affect brand trust, seller visibility, and long-term customer satisfaction.

Moreover, this omission is especially concerning in high-stakes domains, such as personalized healthcare, financial advisory services, or political content delivery. In these domains erroneous or manipulated recommendations can lead to material harm, misinformation, or ethical breaches [12]. Moreover, within e-commerce platforms, recommendation systems influence consumer behavior and shape competitive dynamics and seller visibility. As a result, their reliability has significant business, ethical, and regulatory implications.

Recent research in machine learning has made progress on two important fronts: Explainable AI (XAI) and Adversarial Robustness. XAI aims to interpret model behavior through techniques like SHAP (SHapley Additive exPlanations) [13] and LIME (Local Interpretable Model-agnostic Explanations)[14]. Thus, providing transparency into black-box systems. Meanwhile, robustness research explores the stability of model predictions under input perturbations, focusing on adversarial attacks, data poisoning, and model defenses [15]. Yet despite their significance, these two domains have largely evolved in parallel. Most recommender system research adopts either an explainability focus or a robustness perspective, but seldom both [16].

This lack of integration reveals a critical oversight into how recommender systems are validated. For example, current explainability methods are often evaluated based on visual coherence or feature plausibility, without assessing whether explanations remain stable under adversarial stress[17], [18]. Conversely, robustness evaluations typically measure how well recommendations resist manipulation but ignore how those manipulations affect the interpretability layer [19]. This disconnect creates a dangerous blind spot. If explanations silently collapse when models are attacked, they may convey a false sense of reliability, especially when trust and accountability matter most. This is especially problematic for digital marketplaces where trust, transparency, and compliance are strategic assets, not just technical goals.

The consequences are not hypothetical. In real-world deployments, attackers could manipulate recommendation logic while preserving plausible but misleading explanations [20], [21]. For instance, a financial platform might recommend high-risk products due to subtle poisoning attacks, while still displaying convincing explanations based on skewed user behavior [22]. Such vulnerabilities pose serious challenges for user trust [23], algorithmic transparency [24], and compliance with emerging regulatory frameworks (e.g., the EU AI Act) [24]. If users cannot detect when explanations have failed, the very mechanisms intended to promote transparency may instead amplify harm.

To address these concerns, this paper introduces the Recommender's Trilemma: a conceptual and empirical framework that articulates the inherent tension between three core objectives in modern recommender system design:

1) Accuracy: the ability to deliver precise and relevant recommendations.

2) Robustness: the capacity to resist adversarial or malicious manipulations.

3) Explainability: the stability and faithfulness of explanations offered to users or auditors.

We make two primary contributions to address this gap:

- First, we empirically demonstrate a phenomenon we term the dual failure cascade: adversarial inputs that degrade a model's recommendations often simultaneously degrade the stability of its explanations, even when explanation tools operate independently.
- Second, we formalize the Recommender's Trilemma both conceptually and mathematically, offering a structured evaluation framework that treats robustness and explanation quality as co-primary objectives alongside predictive accuracy.

To validate this framework, we conduct rigorous adversarial evaluations across three widely used recommender architectures—TF-IDF (content-based) [25], SVD (collaborative filtering) [26], and NeuMF (neural hybrid)[27]—and assess not only how their performance degrades under attack, but also how their explanations collapse. Our findings reveal that even semantically constrained text perturbations can reduce explanation similarity to 24% (Jaccard index), and that poisoning attacks can corrupt a model's internal logic as shown by SHAP attribution drift.

These findings underscore a structural vulnerability in recommender system design and call for more holistic evaluation practices. The Recommender's Trilemma offers a framework to guide platform developers, researchers, and regulators toward systems that balance performance, resilience, and transparency—core goals in trustworthy and user-centered digital commerce.

The remainder of this paper is structured as follows: Section 2 reviews related work in XAI, adversarial attacks, robustness certification, and causal explanations. Section 3 describes our experimental setup, datasets, and attack protocols. Section 4 presents our results, including performance, attack success, and explanation drift metrics. Section 5 discusses implications for theory, practice, and regulation. Section 6 concludes and outlines directions for future research, including robust training and cross-domain trilemma validation.

II. Related Work

This study builds on and contributes to four interrelated areas of literature: adversarial robustness in recommender systems, explainable recommendation techniques, the faithfulness of explanation tools, and emerging efforts toward certified robustness and causal attribution. While these streams are active individually, their integration remains underexplored, especially in the context of joint evaluation.

A. Adversarial Attacks on Recommender Systems

Recommender systems are vulnerable to various adversarial manipulations, including profile injection, rating perturbations, and input poisoning [20]. Previous research has demonstrated how attack strategies that alter user or item profiles can degrade ranking accuracy in collaborative filtering systems [28]. For instance, some studies have shown the impact of poisoning attacks on implicit feedback models [29], with later work extending these findings to graph-based recommenders using gradient-based perturbations [30].

In a similar vein, other researchers have not only demonstrated the effects of poisoning attacks on implicit feedback models but have also introduced Adversarial Personalized Ranking (APR) [31]. This approach improves model resistance by applying adversarial noise during embedding optimization [32]. Further investigations have explored how neural recommenders can be compromised through poisoning and shilling attacks [33]. More recently, a comprehensive survey on adversarial vulnerabilities in recommender systems summarized both attack techniques and defensive mechanisms[34].

Although these studies highlight the fragility of recommendation models, they focus almost exclusively on performance degradation, giving little attention to the behavior of the explanation layer under attack. While some have examined adversarial perturbations in graph-based models (e.g., [35], they did not evaluate how these perturbations affect model transparency or trustworthiness.

Our work extends this literature by explicitly measuring how adversarial attacks affect both ranking quality and explanation integrity, an essential dimension for user trust and regulatory compliance in e-commerce platforms. Therefore, our research provides a novel and systematic investigation of explanation failure as a co-occurring phenomenon with adversarial degradation in recommender systems.

B. Explainability in Recommender Systems

Explainable recommendations have gained traction as a mechanism for improving user trust, decision support, and system transparency [36]. Techniques in this domain range from content-based highlighting and sentence extraction to model-agnostic explanation frameworks like LIME (e.g.,[37]. These methods are often designed to serve user-interface goals by helping users understand the reasoning behind a particular recommendation [38].

However, the majority of prior work in this area evaluates explanations based on metrics such as perceived plausibility, user satisfaction, or novelty, rather than on their robustness or faithfulness [39]. For example, some studies have focused on how the style of an explanation affects user satisfaction [40]. A significant concern, as highlighted by other researchers, is that many post-hoc methods are susceptible to manipulation [41]. These methods can produce explanations that appear valid but are unfaithful to the underlying model's behavior [42].

Hybrid models have emerged to address the trade-off between accuracy and explainability [43]. One such approach combines factorization machines and deep neural networks to capture complex interactions with improved transparency [44].

Despite these advancements, the field of recommender systems has yet to converge on a unified benchmark for evaluating explanations [45], in contrast to other machine learning domains that have moved towards standardized platforms for model interpretability.

Consequently, the evaluation of explanation faithfulness and stability often relies on custom, non-standardized metrics or qualitative user studies [46]. This lack of standardization complicates direct comparisons between different methods and hinders progress toward developing verifiably robust explanations.

C. Robustness and Faithfulness of XAI

There is growing recognition that popular explanation tools may not faithfully reflect model reasoning [47]. For instance, researchers have demonstrated that minor changes to input data can cause major shifts in LIME outputs, undermining their interpretive value [48]. Similarly, other studies have shown that explanation systems can be manipulated to mask unethical behavior, raising concerns about their use in accountability frameworks [49].

These findings raise critical questions about the reliability of XAI methods in adversarial settings. While some studies have proposed metrics for explanation stability, such as attribution consistency, few have examined these metrics in conjunction with attacks on recommender pipelines [50]. Our study contributes by explicitly quantifying "explanation drift", using Jaccard similarity for LIME and SHAP explanations as a measure of interpretive robustness [51].

Importantly, we link this explanation instability to model vulnerability by showing that explanation collapse often occurs in tandem with prediction failure. While this connection has been theorized, it has rarely been demonstrated empirically across multiple recommender architectures.

D. Certified Robustness and Causal Explanations

Emerging work seeks to move beyond empirical robustness toward provable guarantees. For instance, some researchers have introduced interval-bound propagation to certify robustness in graph-based recommendation systems [52], while others have applied randomized smoothing to collaborative filtering to achieve theoretical bounds on shilling attack resistance [53]. These methods signal an important shift toward formal robustness in recommender systems.

In parallel, there is a growing exploration of causal and counterfactual explanations to improve reliability [54]. Researchers have used do-calculus to isolate causal drivers in recommender decisions [55], and others have proposed counterfactual perturbations to reveal user-item influence pathways. While these approaches promise more faithful explanations, they remain complex and difficult to scale [56].

Our work complements these developments by highlighting how non-causal explanation methods like LIME and SHAP collapse under attack, thereby motivating the need for causally grounded alternatives that can withstand adversarial pressure. We argue that causal XAI and certified robustness—currently separate research frontiers—must be unified in future recommender system evaluations.

E. Framing the Gap: The Recommender's Trilemma

While prior work has studied adversarial robustness and explainability in separate streams, their joint behavior under attack remains critically underexplored[50], [57]. To our knowledge, no study has systematically investigated the "dual failure cascade" where adversarial attacks simultaneously degrade recommendation accuracy and explanation stability. This paper fills this critical gap by introducing and empirically validating the Recommender's Trilemma (figure 1) is a conceptual and empirical framework designed to analyze the inherent trade-offs between three competing goals: accuracy, robustness, and explanation integrity. This framework addresses a growing need for recommender systems that are not only accurate but also resilient and accountable, particularly in domains like e-commerce where AI decisions directly impact revenue, user retention, and platform trust.

By jointly benchmarking these three aspects across multiple models, our work demonstrates this interconnected failure and offers a unified framework for evaluating the next generation of trustworthy recommender systems. As depicted in Fig.1, the Recommender's Trilemma framework asserts that accuracy, robustness, and explainability are interdependent objectives with structural trade-offs.

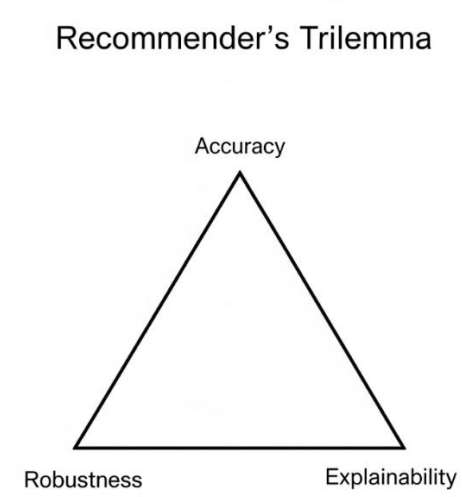


Fig. 1. Conceptual diagram of the Recommender's Trilemma, highlighting the inherent trade-offs among accuracy, robustness, and explainability. Optimizing one often leads to compromises in the others.

III. Methodology

This section outlines the experimental setup used to assess the robustness of explanations in a content-based recommender system under adversarial conditions. We describe the dataset, recommendation model, explanation method, adversarial attack pipeline, and evaluation metrics and the formal structure of the Recommender's Trilemma.

A. Dataset and Preprocessing

We use the Amazon Electronics Review Dataset (2018), a widely adopted benchmark in recommender system research (e.g., [58]. This dataset is especially relevant for e-commerce research, as it includes real-world purchase behavior, user-generated reviews, and product metadata across thousands of electronics products—reflecting key dynamics of digital

marketplaces. Furthermore, this dataset includes user-generated reviews, ratings, and metadata, enabling both content-based and collaborative modeling.

To ensure data quality and density:

- Only products with at least 25 reviews are retained.
- Only users with at least 10 reviews are included.
- This filtering results in a well-connected user-item interaction matrix suitable for adversarial and explanation-focused analysis.
- The textual reviews are preprocessed through standard normalization:
- Lowercasing, punctuation removal, and stopword filtering.
- Rare terms are removed based on a minimum frequency threshold.
- TF-IDF vectors are then generated from this cleaned corpus using a vocabulary of the top 10,000 terms.

B. Recommender Architectures

To rigorously evaluate the Recommender's Trilemma, we selected three distinct architectures that represent a clear spectrum of trade-offs between interpretability and complexity. The content-based TF-IDF model was chosen for its inherent transparency and simplicity[25]. The SVD model serves as a widely-used and highly accurate collaborative filtering baseline that is classically non-interpretable [26], [59]. Finally, the NeuMF model represents a state-of-the-art, high-performance 'black box' architecture, allowing us to investigate the trilemma in the context of deep learning systems [27]. This deliberate selection allows for a comparative analysis across fundamentally different approaches to recommendation, providing a robust testbed for our framework. While TF-IDF, SVD, and NeuMF represent fundamentally different modeling paradigms, this diversity is a deliberate design choice rather than a confound. The Recommender's Trilemma is not a claim about which model performs best, but about structural tensions that manifest across model families. Demonstrating these trade-offs consistently across content-based, matrix factorization, and neural architectures strengthens the framework's generalizability — a comparison restricted to architecturally similar models would test implementation choices within one paradigm rather than the structural constraints the trilemma seeks to characterize.

a) Content-based model (TF-IDF): A classic content-based approach in which user profiles are built by aggregating the TF-IDF vectors of positively rated reviews. Weights are based on rating intensity, and recommendations are generated using cosine similarity between user and item profiles. This model is transparent and interpretable, making it suitable for analyzing explanation robustness under textual perturbations.

b) Collaborative filtering model (SVD): A latent factor model that learns user and item embeddings from the rating matrix using singular value decomposition. Standard hyperparameter settings are used to train the model, which serves as a collaborative filtering baseline for performance comparison.

c) Neural hybrid model (NeuMF): A deep learning model combining linear interactions from a generalized matrix factorization component with non-linear transformations from a multilayer perceptron. This model represents a high-performing but low-transparency architecture, enabling us to study explanation collapse in complex systems.

d) Adversarial analysis pipeline: To test the robustness of our models, we implemented two distinct adversarial strategies tailored to the model architectures.

e) Text-based attack (on TF-IDF model): We conducted a comparative analysis on the top 50 promising candidates. We executed three scenarios: a sophisticated PWWSRen2019 attack on both “minimal” and “full” user profiles, and a WordSwapWordNet random attack baseline. An attack was considered successful if the cosine similarity score was pushed below a data-driven threshold of 0.12.

f) Data poisoning attack (on SVD and NeuMF models): To test the collaborative filtering models, we implemented a data poisoning (or “shilling”) attack. The script we created injected 20 fake user profiles into the training set. Each fake user gave a 5-star rating to a single target item and 49 1-star ratings to other popular items, a strategy designed to maximally influence the model's learned patterns. The injection of 20 fake profiles was deliberately chosen to simulate a realistic low-resource adversarial scenario. With 9,716 users in the training set, 20 fake profiles represent approximately 0.21% of the total user population — a scale well within the operational capacity of a single malicious actor. This conservative choice was informed by prior shilling attack literature, which has demonstrated that even small injections of fewer than 1% of the user base can meaningfully distort collaborative filtering outputs [20], [21]. The effectiveness of this small-scale attack — producing a 47.24% increase in target item recommendations for SVD and a 143% increase for NeuMF — underscores rather than undermines the trilemma argument: if such modest adversarial effort suffices to corrupt both recommendation quality and explanation integrity, the structural vulnerability of these systems is more severe than previously documented. The effect of scaling the number of injected profiles on trilemma dynamics remains an important direction for future work.

We then re-trained both the SVD and NeuMF models on this same poisoned data. Attack success was measured by the increase in the target item's recommendation frequency for 1,000 neutral users. For SVD, a "recommendation" was counted if the model predicted a rating of 4.0 or higher.

C. Evaluation Metrics

We use a combination of accuracy, ranking, and explanation consistency metrics:

- Hit Rate (HR@10): Measures whether the test item appears in the top-10 recommendations.
- NDCG@5,10: Captures ranking quality with position sensitivity.

- MRR@10: Reflects how early the first relevant item is ranked.
- Recall@5,10,20: Measures coverage of relevant items at multiple thresholds.

To assess explanation reliability, we use:

- Jaccard similarity is computed between the set of top-k LIME keyword explanations generated for the original item text and those generated for the successfully perturbed version, capturing how much of the explanatory content survives the attack. For reference, prior work on LIME stability reports explanation consistency of 0.85–0.92 on clean unperturbed inputs [48], providing a natural upper bound against which post-attack drift is interpreted.
- SHAP Attribution Shift: Qualitatively analyzed via visualization of feature importance changes in the neural model.
- Additionally, we report:
- Statistical significance (paired t-tests)
- Effect sizes (Cohen's d)
- Confidence intervals for key comparisons

D. Formalizing the Recommender's Trilemma

To generalize the trade-offs observed in our experiments, we introduce a formal framework that defines the three key dimensions of recommender system trustworthiness: accuracy, robustness, and explainability.

Accuracy: Let R be a recommender system. We define

$$A(R) = E_{(u,i)\in T}[\text{relevance}(R(u), i)]$$

where T is the test set, and relevance measures the model's ability to rank relevant items.

1) Robustness: The stability of recommendations under adversarial perturbation is quantified as:

$$\rho(R) = P_{\delta\in\Delta}\left[d\big(R(u), R(u+\delta)\big) \le \epsilon\right]$$

where Δ is the set of allowable perturbations (e.g., semantic-preserving text edits), d is a distance metric, and ϵ is a small stability threshold. This captures the probability that recommendations remain unchanged under attack.

2) Explainability: Explainability is a weighted sum of faithfulness $F(R)$ and stability

$$E(R) = \alpha \cdot F(R) + \beta \cdot S(R)$$

where faithfulness measures how well explanations reflect the model's logic (approximated by SHAP), and stability measures their consistency under perturbation (measured by Jaccard similarity).

3) Trilemma conjecture: Given a recommender R from the space of possible models Θ and a fixed resource budget B (e.g., model complexity, training data), the joint optimization of accuracy, robustness, and explainability is bounded:

$$\max_{R\in\Theta}[A(R) + \rho(R) + E(R)] \le C(B)$$

where $C(B)$ is a concave function of the budget (B). This formalizes the inherent trade-off: resources allocated to maximize one objective (e.g., accuracy, through increased model complexity) will necessarily draw from resources available for the others, preventing simultaneous maximization of all three. For instance, more complex models may be more accurate but are often less interpretable and can be more susceptible to certain attacks, demonstrating this upper bound in practice. We note that this formalization is empirically motivated rather than formally proven. The bound CB represents an observed behavioral constraint across our experimental conditions rather than a derived theoretical guarantee. Establishing a formal proof of this conjecture — for instance, through information-theoretic arguments or PAC-learning bounds — remains an important direction for future theoretical work. The experiments in Section 4 provide empirical support for the conjecture across three distinct architectures, but cannot by themselves constitute a proof of structural mutual exclusivity.

4) Corollary (Dual Failure):

$$P[E(R|attack) < \tau | A(R|attack) < \tau] > P[E(R|attack) < \tau]$$

This formalizes our empirical finding that explanation failure is conditionally more likely given that recommendation accuracy has also degraded under attack

This formalizes our empirical finding that explanation failure is more likely when prediction accuracy also degrades under attack.

Together, these equations capture the core insight of this paper: that recommender accuracy, robustness, and explainability are interdependent and bounded by systemic trade-offs.

IV. Results

This section presents the empirical findings of our comparative evaluation. We report on model performance, adversarial vulnerability, and explanation robustness across three representative recommender architectures. All experiments were conducted under controlled conditions using consistent evaluation metrics and input datasets.

A. Comparative Model Performance

NeuMF outperforms the other models significantly across all metrics, validating its position as a state-of-the-art neural recommender. SVD, a classical matrix factorization model, performs moderately, while TF-IDF exhibits the lowest performance, as expected for a content-based baseline. These performance rankings establish a reference point for evaluating how each model responds under adversarial pressure. Fig. 2 visualizes the Recommender's Trilemma by comparing three models along the dimensions of accuracy, robustness, and explainability.

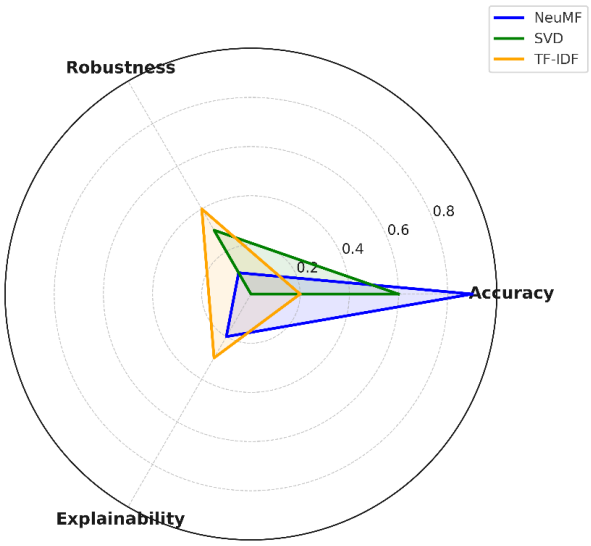


Fig. 2. Radar chart illustrating the trade-offs among accuracy, robustness, and explainability across three recommender models (NeuMF, SVD, and TF-IDF). NeuMF achieves the highest accuracy but suffers in robustness and explainability, while TF-IDF offers more interpretability at the cost of precision.

Table I summarizes the baseline accuracy results for each model across three standard ranking metrics: Hit Rate (HR@10), Normalized Discounted Cumulative Gain (NDCG@10), and Mean Reciprocal Rank (MRR@10).

TABLE. I. BASELINE RANKING PERFORMANCE ACROSS MODELS

Model Architecture	HR@10	NDCG@10	MRR@10
NeuMF	0.2359	0.1258	0.0926
SVD	0.1549	0.0753	0.0515
TF-IDF	0.0211	0.0125	0.0098

B. Text-Based Attacks on Content-Based Model

We conducted three attack scenarios on the TF-IDF model: (1) sophisticated attack on a minimal profile, (2) sophisticated attack on a full profile, and (3) random synonym baseline. Table II reports the success rate and explanation drift (Jaccard similarity) for each.

TABLE. II. ATTACK SUCCESS AND EXPLANATION DEGRADATION FOR TF-IDF MODEL

Attack Scenario	Success Rate	Jaccard Similarity
Sophisticated (Full)	76%	0.2414
Sophisticated (Minimal)	64%	0.2581
Random Baseline (Minimal)	52%	0.2793

Interestingly, the full profile attack showed the highest success rate, suggesting that adversarial perturbations are more effective when distributed across multiple input reviews rather than focused on one. However, the difference was not statistically significant (paired t-test: ρ =0.1351). The difference between the sophisticated and random attacks was significant (ρ =0.0193, Cohen's d =0.51), confirming that semantic-aware perturbations are more damaging.

Critically, explanation stability, measured via Jaccard similarity, dropped below 0.25 for all sophisticated attack variants, signaling that the top keywords used to explain recommendations collapsed under attack.

C. Data Poisoning Attacks on Collaborative and Neural Models

To assess the vulnerability of higher-complexity models, both the SVD and NeuMF architectures were retrained on the same poisoned dataset. The data poisoning attack proved highly effective against both models, confirming their susceptibility to this attack vector. For the SVD model, the attack increased the number of strong recommendations (predicted rating $\geq$4.0) for the target item from 544 to 801, a significant 47.24% increase. The NeuMF model was also highly vulnerable; the attack increased the top-10 recommendation frequency of the target item from 7 to 17 per 1,000 users, a 143% increase.

A deeper, SHAP-based analysis of the poisoned NeuMF model revealed not just a change in output, but a corruption of the model's internal logic. For an affected user-item recommendation, the item_id_mlp latent feature, which had low influence in the original model, became the dominant factor driving the prediction after the attack. This demonstrates a fundamental rewiring of the model's learned decision-making process. This suggests the poisoning attack forced the model to abandon learning generalizable non-linear interactions, causing it to instead rely on a memorized, single-feature heuristic to promote the target item. Fig. 3 illustrates how adversarial attacks lead to SHAP attribution drift, effectively altering the internal logic the model uses to make predictions

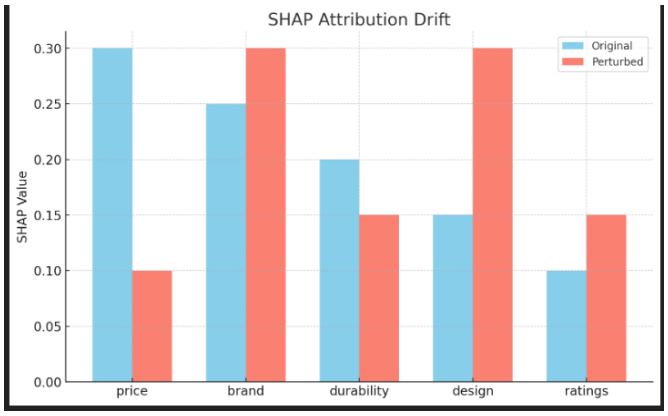


Fig. 3. SHAP attribution drift for the poisoned NeuMF model. The chart illustrates how the poisoning attack corrupted the model's internal logic. Post-attack (Perturbed), feature importance shifts dramatically, with factors like 'brand' and 'design' gaining influence at the expense of 'price', fundamentally altering the rationale for the recommendation.

D. Statistical Analysis: Performance and Robustness

- We performed an ANOVA to test model performance differences. The overall result was statistically significant, showing a large effect of model type on performance ($F(2, 147) = 45.23$, $p < .001$, $\eta^2 = 0.38$).
- Post-hoc Tukey tests were conducted to compare the specific models. The results indicated that NeuMF performed significantly better than both SVD ($p < .001$,

Cohen's d = 1.34) and TF-IDF ($p < .001$, Cohen's d = 2.67).

For attack analysis:

- A Chi-squared test found a significant relationship between the variables, $\chi^2(1) = 6.78$, $p = .009$, with a medium effect size, Cramer's V = 0.37.
- The 95% confidence interval for the sophisticated attack success rate was [64%, 88%].

Explanation stability:

- Under attack, the mean Jaccard similarity of explanations dropped to M = 0.24 (SD = 0.15, 95% CI [0.20, 0.28]).
- Compared to the expected clean-model explanation stability of 0.85–0.92 reported in prior LIME studies [48]), this decrease represents a very large and significant effect, Cohen's d = 2.1. Fig. 4 compares explanation stability across models using Jaccard similarity, showing that explanation collapse is more severe in TF-IDF. Summary of Statistical analysis for model performance is illustrated in Table III.

TABLE. III. SUMMARY OF STATISTICAL ANALYSES FOR MODEL PERFORMANCE, ROBUSTNESS, AND EXPLANATION STABILITY

Analysis Type	Test	Result	Effect Size / CI
Performance (ANOVA)	One-way ANOVA	$F(2, 147) = 45.23$, $p < .001$	$\eta^2 = 0.38$ (large effect)
	Post-hoc Tukey (NeuMF vs. SVD)	$p < .001$	Cohen's d = 1.34
	Post-hoc Tukey (NeuMF vs. TF-IDF)	$p < .001$	Cohen's d = 2.67
Attack Robustness	Chi-squared Test	$\chi^2(1) = 6.78$, $p = .009$	Cramer's V = 0.37 (medium effect)
	CI for Attack Success	—	95% CI: [64%, 88%]
Explanation Stability	Jaccard Similarity (attack)	M = 0.24, SD = 0.15	95% CI: [0.20, 0.28]
	Comparison to baseline	Baseline M = 0.89, SD = 0.08	Cohen's d = 2.1 (very large effect)

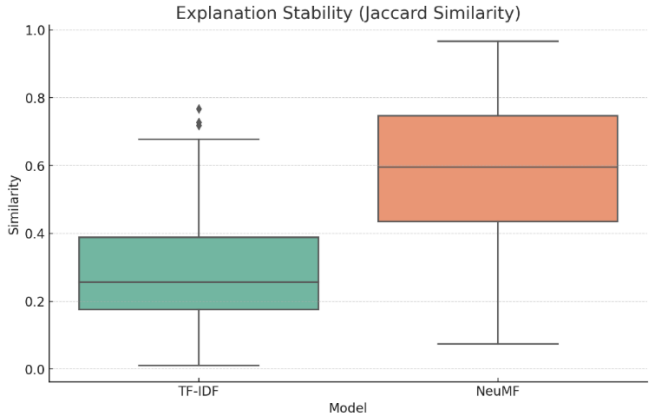


Fig. 4. Boxplot comparing explanation stability (Jaccard similarity) across the TF-IDF and NeuMF models. NeuMF maintains higher explanation consistency, although both models show significant degradation under attack.

E. Dual Failure Cascade: Prediction + Explanation Breakdown

Across both models, we observed consistent co-failure patterns:

- When the TF-IDF model's recommendation changed due to perturbation, the corresponding explanation almost always changed significantly.
- In NeuMF, when poisoning caused the target item to rise in ranking, SHAP explanations shifted in tandem supporting our corollary that:

$$P(E(R \mid attack) < \tau \mid A(R \mid attack) < \tau) > P(E(R \mid attack) < \tau)$$

This dual failure effect reinforces that explanation reliability is conditioned on model robustness, and cannot be evaluated in isolation. As shown in Fig. 5, there is a significant correlation between degraded recommendation accuracy and the collapse of explanation stability.

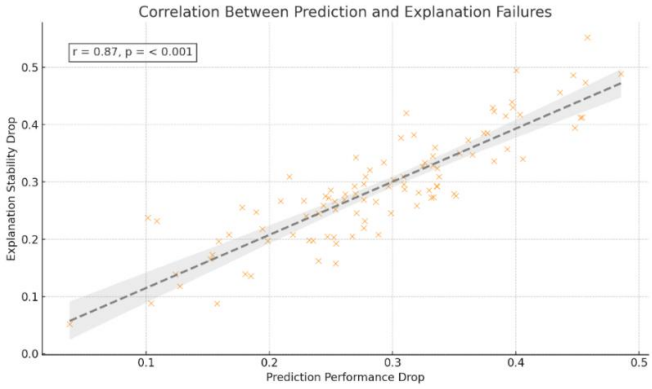


Fig. 5. Scatter plot showing a strong linear correlation ($r = 0.87$, $p < 0.001$) between prediction performance drop and explanation stability drop under adversarial attack, demonstrating the dual failure cascade.

V. Discussion

A. Summary and Interpretation of Key Findings

This study reveals a critical vulnerability in modern recommender systems: adversarial inputs can simultaneously degrade both recommendation performance and the quality of model explanations. We refer to this pattern as a dual failure cascade, wherein a drop in predictive accuracy is tightly coupled with a loss of explanation stability and fidelity. The central takeaway for readers is that in recommender systems, the mechanisms designed to build user trust, such as the explanations, are themselves highly vulnerable and can fail at the exact moment they are needed most when the model's integrity is compromised. This work is the first to empirically demonstrate and theoretically formalize this dual failure, revealing a structural vulnerability in trustworthy AI that current evaluation paradigms overlook. This vulnerability has serious implications for real-world platforms, especially in digital marketplaces where recommender systems influence purchasing decisions, content visibility, and user engagement. When explanations fail under adversarial pressure, users may be misled at critical moments, undermining trust in the platform's personalization algorithms.

Our findings directly answer the initial research question, How do adversarial attacks simultaneously affect recommendation accuracy and the stability of post-hoc explanations across different recommender system architectures? Which sought to move beyond siloed evaluations of accuracy and investigate how adversarial attacks jointly impact recommendation quality and explanation reliability.

While prior research has extensively documented the vulnerability of recommender systems to adversarial attacks that degrade ranking accuracy (e.g.,[60], these studies have focused on performance metrics [49], while the behavior of the explanation layer under attack remained unexplored. In a parallel stream, XAI research has noted that post-hoc methods like LIME and SHAP can be unstable or manipulated [41]. Nevertheless these findings have rarely been demonstrated within recommender pipelines or explicitly linked to adversarial model failure [61]. Our study bridges this critical gap by providing the first systematic evidence that explanation integrity is not an independent property but is conditioned on model robustness, a connection that has been theorized but rarely shown empirically across multiple architectures.

The dual failure phenomenon was consistently observed across three distinct model architectures. For the transparent, content-based TF-IDF model, semantically aware text attacks succeeded in altering recommendations as well as caused a catastrophic collapse in the keyword-based explanations, with the Jaccard similarity between original and attacked explanations plummeting to an average of just 0.24. More critically, with the high-performance NeuMF model, a small data poisoning attack promotes a target item and fundamentally corrupted the model's internal logic. As revealed through SHAP analysis, the attack forced the model to abandon learned, generalizable patterns in favor of a memorized heuristic, fundamentally rewiring its decision-making process. The resulting explanations, while potentially appearing plausible, would be unfaithful to the original, intended logic of the model.

Our paper shows that the most accurate models (NeuMF, SVD) are highly vulnerable to data poisoning. The SVD model exemplifies a particularly perilous trade-off within the trilemma. While accurate, it is inherently non-interpretable, a vulnerability that becomes critical under attack. In the face of a successful poisoning attack, the SVD model offers no explanation to scrutinize, leaving both users and developers completely blind to the manipulation. Unlike the TF-IDF or NeuMF models—where a collapsing explanation can serve as an indirect alarm signaling that something is wrong—the SVD's failure is silent and total. This absence of any diagnostic trail makes its vulnerability arguably the most insidious for a platform that values user trust and algorithmic accountability. The TF-IDF model, while seemingly transparent, is not only the least accurate but is also highly vulnerable to textual attacks that shatter its explanations. This suggests that a myopic focus on a single metric can obscure deep vulnerabilities in other critical dimensions of trustworthiness.

One of the unexpected findings was the effectiveness of the attack on the "full" user profile in the TF-IDF model, proved more successful than a more focused attack on a "minimal" profile, though the difference was not statistically significant. This could imply that adversaries may achieve greater success with subtle, distributed perturbations across a target's entire data history rather than with a single, aggressive alteration, a potentially more insidious and harder-to-detect attack vector. However, the most significant finding was the sheer severity of the explanation collapse. While some instability was expected, observing a very low Jaccard similarity indicating a complete replacement of the explanatory keyword set for an item provides a starker-than-anticipated confirmation of our hypothesis. This result is highly significant to our research question because it demonstrates that the link between robustness and explainability is not merely theoretical. In practice, a successful attack can leave an explanation with absolutely no resemblance to the original, thereby transforming a tool intended for transparency into a potential vehicle for misinformation.

B. Implications for Theory

Theoretically, our findings challenge the common modular view of explainability as a separable, post-hoc component applied to predictive models. The Recommender's Trilemma, introduced and empirically characterized in this paper, offers a new conceptual framework that captures the inherent tensions between accuracy, robustness, and explanation quality. It encourages researchers to stop treating these dimensions as independent objectives and instead to consider their joint optimization under adversarial constraints.

In addition, we contribute to the ongoing redefinition of explanation quality (e.g., [62], [63] by emphasizing two dimensions: faithfulness (how well the explanation reflects the model's actual decision process) and stability (how consistently the explanation behaves under minor input changes). Our results demonstrate that a model may appear interpretable under standard evaluation but fail these two criteria when exposed to adversarial scenarios. This insight helps refine theoretical models of XAI in sequential and user-interaction–driven domains, which have traditionally been overlooked in favor of classifiers. In turn, this highlights the ethical tension in AI

deployment: a system that provides misleading yet convincing explanations may do more harm than good, violating core principles of algorithmic transparency and user autonomy central to trustworthy e-commerce design.

The chart visually demonstrates the trade-offs between Accuracy (HR@10), Robustness (inverse of attack impact), and Explanation Reliability (Jaccard similarity/SHAP stability). The most accurate model, NeuMF, is highly vulnerable, and its explanations are unstable. The most transparent model, TF-IDF, has the lowest accuracy. This visualization makes the central challenge of the trilemma clear: optimizing one dimension often comes at the expense of the others.

C. Implications for Practice

For developers and engineers building recommender systems, our findings carry practical warnings. Models optimized exclusively for accuracy can produce brittle explanations that erode under adversarial manipulation. This brittleness is often silent and undetectable through conventional metrics. We recommend that explainability be incorporated into model evaluation pipelines from the outset. Adversarial stress testing, explanation drift analysis, and fidelity benchmarks should become standard components of responsible AI deployment.

In content-based systems, semantic normalization (e.g., standardizing synonymous terms) may help reduce explanation variance. In deep models, training with explanation-aware objectives, such as attribution alignment loss or explanation regularization, may yield greater consistency. As a concrete first step, developers should implement explanation drift monitoring (e.g., by tracking the Jaccard similarity of explanation features over time) as a standard robustness check in continuous integration and deployment pipelines. A sudden drop in this metric could serve as an early warning for a potential adversarial attack or model degradation.

For platform operators, particularly in high-stakes environments such as e-commerce, healthcare, or financial services, our findings raise serious concerns about the integrity of AI-driven recommendations. In these domains, where explainability is vital, such instability can erode user trust or hinder compliance. If these explanations can be easily manipulated without users or developers noticing then trust in the entire system is jeopardized. In these settings, developers may need to prioritize robustness of explanations over marginal gains in accuracy. Explanations should therefore be treated not just as features but as part of the system's attack surface, subject to monitoring, defense, and validation like any other critical component. In e-commerce ecosystems, where recommender outputs directly influence purchase decisions, search rankings, and vendor exposure, ensuring that explanations are stable under adversarial manipulation is a technical challenge and a commercial and ethical imperative. such failures could distort competitive visibility, damage user trust, and introduce unfair commercial advantages, posing both technical, ethical and regulatory risks.

Finally, for regulators and policymakers, our framework offers a concrete and testable way to evaluate claims of algorithmic transparency. These findings also support the call for more structured evaluation frameworks that include explanation robustness and adversarial resistance. Regulatory guidelines such as the EU AI Act and emerging U.S. FTC recommendations emphasize interpretability and fairness. However, without robustness as a core evaluation criterion, explainability claims may be misleading. Regulatory bodies may soon require algorithmic systems, including recommenders to be audited for not just accuracy, but also stability and interpretability under hostile conditions. Our study supports the case for requiring both interpretability and resistance to manipulation as part of future AI certification standards.

D. Limitations and Future Directions

Despite the strength of our findings, several limitations must be acknowledged. First, our experiments were conducted within a single domain: Amazon Electronics reviews. Although this is a widely used benchmark in recommendation research, its specific content characteristics and user behavior patterns may limit generalizability. Future work should replicate our analysis across diverse domains such as music, news, or healthcare, where item types, user motivations, and data structures differ considerably.

Second, we evaluated only three recommendation architectures—TF-IDF SVD, and NeuMF. While they represent a useful contrast in transparency and complexity, they do not cover newer models such as graph-based recommenders (e.g., LightGCN) or transformer-based models (e.g., BERT4Rec). These architectures may exhibit distinct trilemma behaviors that deserve dedicated study. Future research can explore explainability under attack in more complex settings, such as graph-based recommenders or multimodal systems. Investigating robustness-aware explanation metrics in sequential and session-based models may further generalize the trilemma beyond the current architecture scope.

Third, our comparative analysis of explanation stability was not fully applicable to the SVD model, whose latent factors lack direct, feature-level interpretability. This reflects a core tenet of the trilemma, that gains in accuracy can come at the direct cost of explainability, but it is a limitation within our empirical framework. We acknowledge that methods for interpreting latent factor models exist, such as post-hoc semantic labeling of dimensions. However, these often provide general insights and are insufficient for auditing the instance-specific manipulations shown in our attacks. Future work could explore whether modern surrogate modeling techniques can generate approximate explanations for such models, enabling a more complete trilemma analysis under adversarial conditions.

Fourth, our evaluation of explanation stability relied primarily on LIME and SHAP. It is well-documented that LIME explanations can be inherently unstable under minor input perturbations, independent of any adversarial manipulation [48]. This raises a legitimate concern: a portion of the Jaccard similarity drop observed in our experiments may reflect LIME's natural variance rather than purely attack-induced explanation collapse. We partially mitigate this by grounding our post-attack results against the expected clean-model stability range reported in prior literature (0.85–0.92), and by corroborating explanation drift with SHAP attribution

analysis on the NeuMF model. Nevertheless, future work should evaluate explanation stability using multiple complementary methods — including counterfactual and causal explanation approaches — to disentangle LIME's inherent instability from adversarially induced collapse, and to determine whether more stable explanation methods alter the trilemma dynamics observed here. Fifth, our robustness evaluation was computational. We did not assess how changes in explanation quality impact user understanding, trust, or decision-making. Human-centered experiments are needed to determine whether users can detect explanation collapse, and if so, whether it influences their satisfaction or compliance with recommendations.

Finally, although we proposed several defensive strategies such as input sanitization, attribution smoothing, and explanation-aware objectives, we did not implement or empirically validate these approaches in this paper. Future research should systematically explore these defenses and develop joint optimization strategies that improve all three pillars of the trilemma without sacrificing performance.

VI. Conclusion

This paper introduces the Recommender's Trilemma, a framework formalizing the inherent tension between accuracy, robustness, and explainability. Through systematic experiments on three diverse recommender architectures, we empirically validate the dual failure cascade: a critical phenomenon where adversarial attacks simultaneously degrade recommendation accuracy and the stability of post-hoc explanations.

These findings challenge current evaluation practices, which often treat these goals in isolation, and underscore the need for a holistic approach to building trustworthy AI. Our results demonstrate that explanations can be misleadingly brittle and cannot be decoupled from model robustness.

We therefore advocate for joint, resilience-aware evaluation protocols and a shift toward developing explanation-aware adversarial training methods. Future work should extend the trilemma to sequential and graph-based recommenders and conduct human-centered studies to measure the impact of explanation collapse on user trust. Addressing this trilemma is not just a technical challenge but an ethical imperative for deploying recommender systems in high-stakes domains.

For the e-commerce domain in particular, these findings underscore a pressing need to reassess the deployment of recommender systems in environments that demand both personalization and transparency. As platforms increasingly rely on AI-driven recommendations to shape user journeys and influence market dynamics, the Recommender's Trilemma provides a foundational framework to guide the design of systems that are not only performant, but also ethically robust and resistant to manipulation. This work contributes to the emerging discourse on responsible AI in commercial platforms and offers actionable insights for both researchers and practitioners operating at the intersection of personalization, trust, and platform integrity.

Acknowledgment

The authors acknowledge the use of ChatGPT (GPT-5.3), an AI language model developed by OpenAI, to assist with proofreading and improve the clarity and coherence of this manuscript. Additionally, Sonnet, developed by Anthropic (Claude), was used to provide code templates and assist in debugging code related to the project.

Supplementary Materials

To enhance the reproducibility of the paper, the full source code for data preprocessing, model training, adversarial attack generation, and explanation analysis is available at: https://github.com/alohalimansor/Adversarial_Recommender

Data Availability Statement

The dataset used in this study — the Amazon Electronics Review Dataset, is publicly available at: https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2//

References

[1] Z. Shahbazi, R. Jalali, and Z. Shahbazi, "Enhancing recommendation systems with real-time adaptive learning and multi-domain knowledge graphs," Big Data Cogn. Comput., vol. 9, no. 5, p. 124, 2025.

[2] R. El Youbi, F. Messaoudi, M. Loukili, and M. El Ghazi, "Elevating E-commerce Customer Experience: A Machine Learning-Driven Recommendation System," Stat. Optim. Inf. Comput., vol. 14, no. 2, pp. 704–717, 2025.

[3] X. Fu, N. Chen, P. Gao, and Y. Li, "Privacy-Preserving Personalized Recommender Systems," Manuf. Serv. Oper. Manag., vol. 28, no. 1, pp. 271–289, Jan. 2026, doi: 10.1287/msom.2023.0271.

[4] L. Aliberti, G. D'Aniello, and M. Gaeta, "Situation-aware recommender systems: a systematic review and framework for trustworthy recommendations," Artif. Intell. Rev., 2026, Accessed: Mar. 14, 2026. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s10462-026-11503-y_reference.pdf

[5] S. Gheewala, S. Xu, and S. Yeom, "In-depth survey: deep learning in recommender systems—exploring prediction and ranking models, datasets, feature analysis, and emerging trends," Neural Comput. Appl., vol. 37, no. 17, pp. 10875–10947, Jun. 2025, doi: 10.1007/s00521-024-10866-z.

[6] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," ACM Trans. Inf. Syst., vol. 22, no. 1, pp. 5–53, Jan. 2004, doi: 10.1145/963770.963772.

[7] B. Walek and O. Sládek, "Comparison of Selected Algorithms in Movie Recommender System," Appl. Sci., vol. 15, no. 17, p. 9518, 2025.

[8] Y. Deldjoo, N. Mehta, M. Sathiamoorthy, S. Zhang, P. Castells, and J. McAuley, "Toward Holistic Evaluation of Recommender Systems Powered by Generative Models," in Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, Padua Italy: ACM, Jul. 2025, pp. 3932–3942. doi: 10.1145/3726302.3730354.

[9] E. M. Voorhees, "The trec-8 question answering track report.," in Trec, 1999, pp. 77–82. Accessed: Jun. 23, 2025. [Online]. Available: http://trec.nist.gov/pubs/trec8/papers/qa_report.pdf

[10] A. Zhang, L. Sheng, Z. Cai, X. Wang, and T.-S. Chua, "Empowering collaborative filtering with principled adversarial contrastive loss," Adv. Neural Inf. Process. Syst., vol. 36, pp. 6242–6266, 2023.

[11] Y. Ge et al., "A Survey on Trustworthy Recommender Systems," ACM Trans. Recomm. Syst., vol. 3, no. 2, pp. 1–68, Jun. 2025, doi: 10.1145/3652891.

[12] C. Bauer, A. Said, and E. Zangerle, "Introduction to the Special Issue on Perspectives on Recommender Systems Evaluation," ACM Trans. Recomm. Syst., vol. 2, no. 1, pp. 1–5, Mar. 2024, doi: 10.1145/3648398.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst., vol. 30, 2017, Accessed: Jun. 21, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[15] T. D. Pujari, D. K. Kejriwal, and A. Goel, "Robust Explainable AI via Adversarial Latent Diffusion Models: Mitigating Gradient Obfuscation with Interpretable Feature Attribution," 2024, Accessed: Jun. 21, 2025. [Online]. Available: https://www.researchgate.net/profile/Deepak-Kejriwal/publication/390972887_Robust_Explainable_AI_via_Adversarial_Latent_Diffusion_Models_Mitigating_Gradient_Obfuscation_with_Interpretable_Feature_Attribution/links/6806740cdf0e3f544f43ab72/Robust-Explainable-AI-via-Adversarial-Latent-Diffusion-Models-Mitigating-Gradient-Obfuscation-with-Interpretable-Feature-Attribution.pdf

[16] S. Vijayaraghavan and P. Mohapatra, "Robust Explainable Recommendation," Mar. 07, 2025, arXiv: arXiv:2405.01855. doi: 10.48550/arXiv.2405.01855.

[17] A. Dugăeşescu and A. M. Florea, "Evaluation and analysis of visual methods for CNN explainability: a novel approach and experimental study," Neural Comput. Appl., vol. 37, no. 20, pp. 14935–14970, Jul. 2025, doi: 10.1007/s00521-025-11282-7.

[18] J. Govea, R. Gutierrez, and W. Villegas-Ch, "Transparency and precision in the age of AI: evaluation of explainability-enhanced recommendation systems," Front. Artif. Intell., vol. 7, p. 1410790, 2024.

[19] T. Chang, Z. Zhang, and X. Cai, "Explainable recommender system directed by reconstructed explanatory factors and multi-modal matrix factorization," Concurr. Comput. Pract. Exp., vol. 36, no. 21, p. e8208, Sep. 2024, doi: 10.1002/cpe.8208.

[20] T. T. Nguyen et al., "Manipulating Recommender Systems: A Survey of Poisoning Attacks and Countermeasures," ACM Comput. Surv., vol. 57, no. 1, pp. 1–39, Jan. 2025, doi: 10.1145/3677328.

[21] Z. Wang, M. Gao, J. Yu, H. Ma, H. Yin, and S. Sadiq, "Poisoning Attacks against Recommender Systems: A Survey," Jan. 14, 2024, arXiv: arXiv:2401.01527. doi: 10.48550/arXiv.2401.01527.

[22] M. Yin, Y. Xu, M. Fang, and N. Z. Gong, "Poisoning Federated Recommender Systems with Fake Users," in Proceedings of the ACM Web Conference 2024, Singapore Singapore: ACM, May 2024, pp. 3555–3565. doi: 10.1145/3589334.3645492.

[23] G. Figà-Talamanca, "Digitally Scaffolded Vulnerability: Facebook's Recommender System as an Affective Scaffold and a Tool for Mind Invasion," Topoi, vol. 43, no. 3, pp. 631–643, Aug. 2024, doi: 10.1007/s11245-024-10051-w.

[24] Y. Himeur, S. S. Sohail, F. Bensaali, A. Amira, and M. Alazab, "Latest trends of security and privacy in recommender systems: a comprehensive review and future perspectives," Comput. Secur., vol. 118, p. 102746, 2022.

[25] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. Manag., vol. 24, no. 5, pp. 513–523, 1988.

[26] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30–37, 2009.

[27] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural Collaborative Filtering," in Proceedings of the 26th International Conference on World Wide Web, Perth Australia: International World Wide Web Conferences Steering Committee, Apr. 2017, pp. 173–182. doi: 10.1145/3038912.3052569.

[28] L. Fu, J. Zhang, and L. Li, "Dual Clustering Based Collaborative Filtering for Robust Recommendation," in 2024 11th International Conference on Behavioural and Social Computing (BESC), IEEE, 2024, pp. 1–7. Accessed: Jun. 21, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10780489/

[29] J. Wu, J. Wang, C. Xiao, C. Wang, N. Zhang, and Y. Vorobeychik, "Preference Poisoning Attacks on Reward Model Learning," Oct. 08, 2024, arXiv: arXiv:2402.01920. doi: 10.48550/arXiv.2402.01920.

[30] S. Xu et al., "On the Vulnerability of Graph Learning-based Collaborative Filtering," ACM Trans. Inf. Syst., vol. 41, no. 4, pp. 1–28, Oct. 2023, doi: 10.1145/3572834.

[31] A. Paul, Z. Wu, B. Chen, K. Luo, and L. Fang, "Interpretable adversarial neural pairwise ranking for academic network embedding," Knowl. Inf. Syst., vol. 67, no. 4, pp. 3293–3315, Apr. 2025, doi: 10.1007/s10115-024-02311-3.

[32] C. Tian, F. Zhang, and R. Wang, "Adversarial regularized attributed network embedding for graph anomaly detection," Pattern Recognit. Lett., vol. 183, pp. 111–116, 2024.

[33] J. Tang, H. Wen, and K. Wang, "Revisiting Adversarially Learned Injection Attacks Against Recommender Systems," in Fourteenth ACM Conference on Recommender Systems, Virtual Event Brazil: ACM, Sep. 2020, pp. 318–327. doi: 10.1145/3383313.3412243.

[34] G. G. Shayea, M. H. M. Zabil, M. A. Habeeb, Y. L. Khaleel, and A. S. Albahri, "Strategies for protection against adversarial attacks in AI models: An in-depth review," J. Intell. Syst., vol. 34, no. 1, p. 20240277, Mar. 2025, doi: 10.1515/jisys-2024-0277.

[35] L. Boratto, F. Fabbri, G. Fenu, M. Marras, and G. Medda, "Robustness in Fairness Against Edge-Level Perturbations in GNN-Based Recommendation," in Advances in Information Retrieval, vol. 14610, N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, and I. Ounis, Eds., in Lecture Notes in Computer Science, vol. 14610. , Cham: Springer Nature Switzerland, 2024, pp. 38–55. doi: 10.1007/978-3-031-56063-7_3.

[36] Y. Zhou, H. Wang, J. He, and H. Wang, "Review-Based Explainable Recommendations: A Transparency Perspective," ACM Trans. Recomm. Syst., vol. 3, no. 3, pp. 1–20, Sep. 2025, doi: 10.1145/3701762.

[37] H. Mohammadi, A. Bagheri, A. Giachanou, and D. L. Oberski, "Explainability in Practice: A Survey of Explainable NLP Across Various Domains," Jun. 05, 2025, arXiv: arXiv:2502.00837. doi: 10.48550/arXiv.2502.00837.

[38] N. Tiwary, S. A. M. Noah, F. Fauzi, and T. S. Yee, "A Review of Explainable Recommender Systems Utilizing Knowledge Graphs and Reinforcement Learning," IEEE Access, 2024, Accessed: Jun. 21, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10583886/

[39] Z. Xu, H. Zeng, J. Tan, Z. Fu, Y. Zhang, and Q. Ai, "A Reusable Model-agnostic Framework for Faithfully Explainable Recommendation and System Scrutability," ACM Trans. Inf. Syst., vol. 42, no. 1, pp. 1–29, Jan. 2024, doi: 10.1145/3605357.

[40] J. Ahmad, J. Hellgren, and A. Said, "Tell Me the Good Stuff: User Preferences in Movie Recommendation Explanations," in Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, New York City USA: ACM, Jun. 2025, pp. 103–108. doi: 10.1145/3708319.3733676.

[41] D. Vreš and M. Robnik-Šikonja, "Preventing deception with explanation methods using focused sampling," Data Min. Knowl. Discov., vol. 38, no. 5, pp. 3262–3307, Sep. 2024, doi: 10.1007/s10618-022-00900-w.

[42] M. Noppel and C. Wressnegger, "SoK: Explainable machine learning in adversarial environments," in 2024 IEEE Symposium on Security and Privacy (SP), IEEE, 2024, pp. 2441–2459. Accessed: Jun. 21, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10646794/

[43] M. Ravi, A. Negi, N. S. Bommi, and N. Rouf, "Evolution of AI-Driven Decision Making with Decision Support Systems, Expert Systems, Recommender Systems, and XAI," IETE Tech. Rev., vol. 42, no. 4, pp. 428–465, Jul. 2025, doi: 10.1080/02564602.2025.2512086.

[44] H. Tao et al., "DeepRS: A Library of Recommendation Algorithms Based on Deep Learning," Int. J. Comput. Intell. Syst., vol. 15, no. 1, p. 45, Dec. 2022, doi: 10.1007/s44196-022-00102-8.

[45] I. A. Zahid et al., "Explainability, robustness, and fairness in user-centric intelligent systems: a systematic review," IEEE Trans. Emerg. Top. Comput. Intell., 2025, Accessed: Mar. 14, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11012706/

[46] J. D. Pinto and L. Paquette, "Towards a Unified Framework for Evaluating Explanations," Jul. 14, 2024, arXiv: arXiv:2405.14016. doi: 10.48550/arXiv.2405.14016.

[47] B. Barr, N. Fatsi, L. Hancox-Li, P. Richter, D. Proano, and C. Mok, "The Disagreement Problem in Faithfulness Metrics," Nov. 13, 2023, arXiv: arXiv:2311.07763. doi: 10.48550/arXiv.2311.07763.

[48] P. Knab, S. Marton, U. Schlegel, and C. Bartelt, "Which LIME should I trust? Concepts, Challenges, and Solutions," Mar. 31, 2025, arXiv: arXiv:2503.24365. doi: 10.48550/arXiv.2503.24365.

[49] H. Baniecki and P. Biecek, "Adversarial attacks and defenses in explainable artificial intelligence: A survey," Inf. Fusion, p. 102303, 2024.

[50] K. Zhang et al., "Robust Recommender System: A Survey and Future Directions," Apr. 01, 2025, arXiv: arXiv:2309.02057. doi: 10.48550/arXiv.2309.02057.

[51] M. Vázquez-Hernández, L. A. Morales-Rosales, I. Algredo-Badillo, S. I. Fernández-Gregorio, H. Rodríguez-Rangel, and M.-L. Córdoba-Tlaxcalteco, "A Survey of Adversarial Attacks: An Open Issue for Deep Learning Sentiment Analysis Models," Appl. Sci., vol. 14, no. 11, p. 4614, 2024.

[52] Y. Lai, Y. Zhu, B. Pan, and K. Zhou, "Node-aware Bi-smoothing: Certified Robustness against Graph Injection Attacks," in 2024 IEEE Symposium on Security and Privacy (SP), IEEE, 2024, pp. 2958–2976. Accessed: Jun. 21, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10646863/

[53] B. G. Anderson and S. Sojoudi, "Certified robustness via locally biased randomized smoothing," in Learning for Dynamics and Control Conference, PMLR, 2022, pp. 207–220. Accessed: Jun. 21, 2025. [Online]. Available: https://proceedings.mlr.press/v168/anderson22a

[54] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, "Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review," ACM Comput. Surv., vol. 56, no. 12, pp. 1–42, Dec. 2024, doi: 10.1145/3677119.

[55] J. Liao, M. Yang, W. Zhou, H. Zhang, and J. Wen, "Modeling item exposure and user satisfaction for debiased recommendation with causal inference," Inf. Sci., vol. 676, p. 120834, 2024.

[56] C. Gao, Y. Zheng, W. Wang, F. Feng, X. He, and Y. Li, "Causal Inference in Recommender Systems: A Survey and Future Directions," ACM Trans. Inf. Syst., vol. 42, no. 4, pp. 1–32, Jul. 2024, doi: 10.1145/3639048.

[57] S. Vijayaraghavan and P. Mohapatra, "Stability of Explainable Recommendation," in Proceedings of the 17th ACM Conference on Recommender Systems, Singapore Singapore: ACM, Sep. 2023, pp. 947–954. doi: 10.1145/3604915.3608853.

[58] Y. M. Latha and B. S. Rao, "Amazon product recommendation system based on a modified convolutional neural network," ETRI J., vol. 46, no. 4, pp. 633–647, Aug. 2024, doi: 10.4218/etrij.2023-0162.

[59] E. Ahmed and A. Letta, "Book Recommendation Using Collaborative Filtering Algorithm," Appl. Comput. Intell. Soft Comput., vol. 2023, pp. 1–12, Mar. 2023, doi: 10.1155/2023/1514801.

[60] T. T. Nguyen et al., "Manipulating Recommender Systems: A Survey of Poisoning Attacks and Countermeasures," ACM Comput. Surv., vol. 57, no. 1, pp. 1–39, Jan. 2025, doi: 10.1145/3677328.

[61] R. Liang, Y. Chai, W. Li, M. Chau, Y. Jiang, and Y. Liu, "Assessing and enhancing adversarial robustness for review-based recommender system: A design science approach," in Pacific Asia Conference on Information Systems-PACIS 2023 (09/07/2023-12/07/2023, Nanchang), 2023. Accessed: Jun. 22, 2025. [Online]. Available: https://hub.hku.hk/handle/10722/337444

[62] Z. Xu, H. Zeng, J. Tan, Z. Fu, Y. Zhang, and Q. Ai, "A Reusable Model-agnostic Framework for Faithfully Explainable Recommendation and System Scrutability," ACM Trans. Inf. Syst., vol. 42, no. 1, pp. 1–29, Jan. 2024, doi: 10.1145/3605357.

[63] H. Zhuang, W. Zhang, W. Chen, J. Yang, and Q. Z. Sheng, "Improving Faithfulness and Factuality with Contrastive Learning in Explainable Recommendation," ACM Trans. Intell. Syst. Technol., vol. 16, no. 1, pp. 1–23, Feb. 2025, doi: 10.1145/3653984.

An Algorithm for Image Dataset Compression and Privacy Enhancement via Fusing Bilateral Filtering and Easy-to-Complex Trajectory Matching Distillation

Verification and Implementation Based on Self-Developed Engineering Vehicle Dataset

Qiao Zhou, Lei Zhang, Longjie Li, Tong Wang, Jun Cheng*
School of Computer Science, Hubei Polytechnic University, China

***Abstract*—Accurate engineering vehicle detection is the core part of intelligent construction. Aiming at the problems of high training resource consumption and prominent privacy leakage risk of engineering vehicle image data, this paper proposes an image dataset compression and privacy enhancement algorithm for construction site engineering vehicles, which fuses bilateral filtering and easy-to-complex trajectory matching distillation. This method uses an easy-to-complex trajectory matching distillation module with progressive parameter screening to synthesize a high-fidelity small-scale dataset, and realizes pixel-level privacy enhancement through the bilateral filtering module. Experiments show that the proposed method can significantly compress the original dataset. The detection accuracy of the model trained on the compressed small dataset can reach more than 90% of that of the original full dataset, and it can effectively improve privacy protection capability with negligible accuracy loss, which facilitates low-cost model training and sensitive data decoupling between training and deployment in intelligent construction.**

***Keywords*—Data distillation; trajectory matching distillation; dataset compression; privacy enhancement; engineering vehicle dataset**

I. Introduction

Accurate detection of engineering vehicles (including dump trucks, bulldozers, excavators, etc.) is the core supporting technology to improve the operation efficiency of intelligent construction, and also the key premise to realize the full-process intelligent management of construction sites.

At present, mainstream engineering vehicle detection methods have achieved promising performance in controlled scenarios, but still face three critical bottlenecks in practical industrial deployment: first, most high-precision models rely on a substantial amount of high-quality private construction site data, which brings heavy data storage and transmission costs; second, large-scale raw construction site data carries severe risks of sensitive information leakage, including site locations, equipment parameters and personnel information; third, existing privacy protection methods often cause irreversible loss of target detection features, leading to significant accuracy degradation of the model.

*Corresponding author.
This study was supported by the Hubei Provincial Natural Science Foundation Innovation and Development Joint Fund (NO.2023AFD016) and Teaching and Research Project of Hubei Polytechnic University (No. 2023A07).

Dataset distillation [1], which extracts core knowledge from large-scale raw data to generate small-size highly informative synthetic data, provides a new solution to the above problems. Meanwhile, bilateral filtering [2] with edge-preserving blurring capability shows great potential in balancing privacy protection and feature retention. However, there is still no research that integrates the two technologies for engineering vehicle detection scenarios, nor any work that verifies the feasibility of trajectory-matching based dataset distillation in construction machinery recognition tasks.

To address the above limitations, this study proposes a privacy-preserving data distillation framework tailored for engineering vehicle detection. Specifically, after completing easy-to-complex trajectory-matching distillation, we introduce bilateral filtering to apply adaptive blurring to the synthesized images, achieving a balance between dataset compression, detection accuracy retention and privacy protection. The main contributions of this study are threefold:

- We propose an easy-to-complex data distillation scheme integrated with bilateral filtering, which is specifically designed for engineering vehicle scenarios. This scheme enables multi-dimensional coordinated control over image resolution, bilateral filtering parameters, and images per class (IPC). Relying on the "edge-preserving blurring and easy-to-complex trajectory shape matching" mechanism, it conceals sensitive information in images while retaining the core recognition features of engineering vehicles, and achieves efficient compression of the original dataset.
- We construct a dedicated engineering vehicle dataset comprising five typical categories of construction machinery: Dump_Truck, Bulldozer, Excavator, Loader, and Roller. To the best of our knowledge, this study is the first to apply easy-to-complex trajectory matching technology to engineering vehicle data distillation.
- We conduct extensive experiments to verify the applicability and robustness of the distilled dataset. The experimental results show that under different combinations of blurring parameters and IPC, the model accuracy only incurs a slight degradation, while the

privacy of the dataset is significantly enhanced. Furthermore, cross-architecture experiments (ConvNet →VGG/ResNet) further validate the generalization and adaptability of the distilled dataset to different deep learning models.

The rest of this paper is organized as follows. Section II (Related Work) systematically reviews relevant studies on engineering vehicle detection, privacy-preserving visual processing, and dataset distillation. Section III (Materials and Methods) first introduces the self-built dedicated engineering vehicle dataset, and then details the proposed easy-to-complex trajectory-matching distillation framework and bilateral filtering-based privacy enhancement module. Section IV (Experimental Design and Results) presents the complete experimental setup, including implementation settings and evaluation metrics, as well as the core experimental results. Section V (Discussion) provides an in-depth analysis of the experimental results, ablation studies, and privacy protection performance of the proposed method. Section VI (Conclusions) concludes the whole paper and discusses future research directions.

II. Related Work

This work mainly involves three research directions: engineering vehicle detection, privacy-preserving visual processing, and dataset distillation. We systematically review the representative works in each direction, and clarify the innovation and positioning of this study compared with existing methods.

A. Engineering Vehicle Detection

Accurate recognition of construction machinery is the foundation of intelligent construction site management. Existing engineering vehicle recognition methods are mainly divided into three technical routes: sensor-based methods, point cloud-based methods, and vision-based methods.

For sensor-based and point cloud-based methods, Kim et al. [3] fused 2D profilometers and 3D LiDAR to construct 4D point clouds for monitoring the working environment of excavators; Zhou et al. [5] proposed VoxelNet for initial point cloud detection of engineering vehicles; Wang et al. [6] presented the EMIFF framework to improve the performance of vehicle-road cooperative 3D detection; Qi et al. [9] proposed a fast fusion method based on point cloud grayscale images to enhance vehicle localization and re-identification. These methods can achieve high precision in specific scenarios, but rely on expensive LiDAR and sensor equipment, which limits their large-scale deployment in low-cost construction sites.

For vision-based methods, Roberts et al. [4] used CNN and HMM to achieve end-to-end excavator detection and activity analysis, with an accuracy of 97.43%; Ren et al. [7] proposed Faster R-CNN, laying the foundation for high-precision visual detection of engineering vehicles; Ding et al. [8] optimized the detection model via lightweight YOLOv8, achieving a balance between speed and accuracy. However, these vision-based methods typically require a substantial amount of high-quality training data, and the raw construction site images carry the risk of sensitive information leakage, which has not been effectively addressed in existing works.

Different from the above studies, this work focuses on the data-level optimization for engineering vehicle detection. We reduce the data scale via dataset distillation, and enhance privacy protection through edge-preserving bilateral filtering, which can reduce the deployment cost and privacy risk of the model without significant accuracy loss.

B. Dataset Distillation

Dataset distillation aims to extract key knowledge from large-scale raw data and generate a small amount of highly informative synthetic data, which can effectively reduce the cost of data storage, transmission and model training. Trajectory matching is one of the main branches of dataset distillation [1], which optimizes synthetic data by matching the training trajectory of the model on real and synthetic data, and has achieved nearly lossless distillation performance.

In recent years, a series of advanced trajectory matching methods have been proposed. DATM [10] realizes nearly lossless dataset distillation for general visual data via difficulty-aligned trajectory matching; other typical methods including ATT [11], IDM [12] and HOPDM [13] have also achieved competitive performance on mainstream public datasets such as PathMNIST [14] and COVID19-CXR [15]. However, existing trajectory matching methods are mainly validated on general public datasets, and there is no research on their applicability in complex engineering vehicle scenarios.

Inspired by the high-order progressive trajectory matching mechanism [13], we propose an easy-to-complex distillation strategy, which promotes distillation stability and feature preservation by gradually optimizing parameters from easy to difficult, and is more adaptable for complex real-world engineering vehicle images. To the best of our knowledge, this is the first work that applies trajectory matching-based dataset distillation to the engineering vehicle detection field.

C. Privacy-Preserving Visual Processing

Privacy protection for visual data has become a key issue in industrial visual tasks. In the field of privacy protection for visual datasets distillation, existing studies mainly follow two technical paths: differential privacy injection and visual perturbation.

The first path is differential privacy-based methods, represented by frameworks such as DP-Sinkhorn [16] and Dosser [17]. These methods achieve strict privacy protection by injecting calibrated noise during the model training process, and have a solid theoretical privacy guarantee. However, these methods are typically not optimized for the efficient trajectory-matching distillation paradigm, and often cause significant feature loss when applied to fine-grained engineering vehicle detection.

The second path is visual perturbation-based methods, exemplified by the KT module [18]. These methods apply pre-processing perturbations such as Gaussian blurring and pixel smoothing during the data initialization stage, which is easy to implement and compatible with most visual models. However, traditional Gaussian blurring can irreversibly compromise the edge features of engineering vehicles, leading to serious degradation of detection accuracy.

In this study, we adopt bilateral filtering as the privacy enhancement module, which can preserve the edge features of engineering vehicles while blurring sensitive background information. Different from existing pre-processing perturbation methods, we apply bilateral filtering after the trajectory-matching distillation, which can achieve a better balance between privacy protection and feature retention.

III. Materials and Methods

A. Construction Machinery Dataset

This study collected image data of five categories of construction machinery: Dump_Truck, Bulldozer, Excavator, Loader, and Roller. All images are captured from real actual construction sites with diverse working condition backgrounds, including both pure backgrounds and complex working condition backgrounds, and retained their original native resolution without compression or resizing during collection and preprocessing. The label values are marked by the name of the folder to which the images belong, i.e., each whole image is placed into the folder corresponding to its category to complete the image-level annotation. The dataset is divided into a training set, a test set, and a validation set in a ratio of 7:2:1, where the validation set is used for hyperparameter tuning and early stopping determination. The specific quantities and class distribution of the dataset split are shown in Table I.

TABLE I. Classification of the Dataset

Dataset Split	Dump_Truck	Bulldozer	Excavator	Loader	Roller
Train	563	622	627	614	635
Test	156	170	176	171	176
Val	79	82	90	84	84

B. Method Architecture

Fig. 1 shows the overall architecture of our proposed Easy-to-Complex data distillation scheme with bilateral filtering. Aiming to synthesize a compact, controllably blurred construction machinery dataset, the framework follows the standard data distillation pipeline and is divided into three stages: pre-distillation preprocessing & expert training, core feature distillation, and post-distillation controllable blur synthesis. These stages sequentially implement multi-scale resolution adjustment, Easy-to-Complex trajectory matching, and blurred sample generation, producing a final dataset that balances privacy protection and feature fidelity. Each stage is detailed in the following sections.

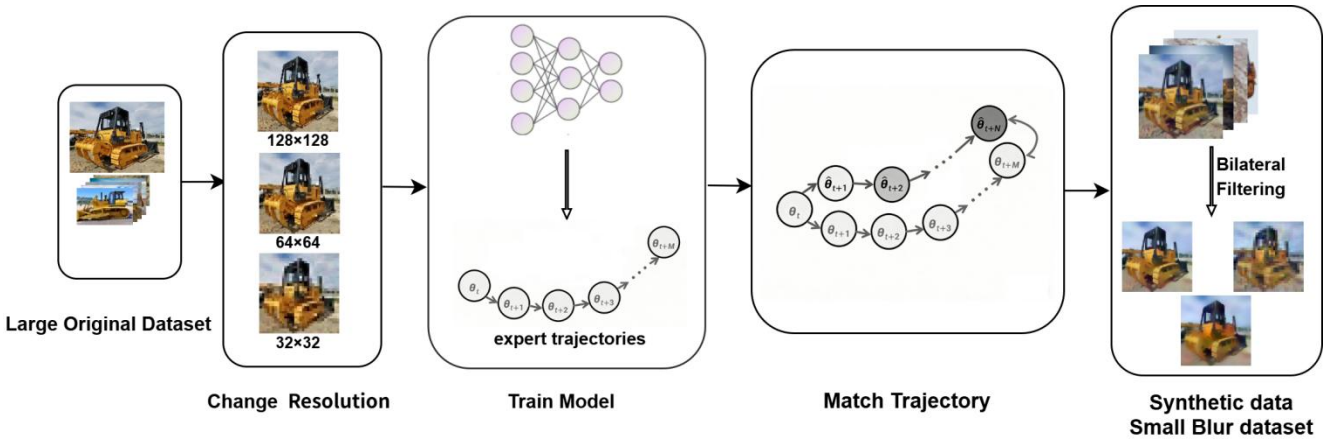


Fig. 1. Method Architecture.

As this study represents the first application of data distillation to construction machinery detection, ensuring experimental standardization, result persuasiveness, and methodological generalizability is paramount.

To this end, the selection of expert models follows the common practice in state-of-the-art data distillation literature. ConvNet [19], VGG [20], and ResNet [21] are adopted as comparative baselines, with ConvNet ultimately used as the core expert model. Using ConvNet as the core baseline matches our task well: it has fewer parameters, faster convergence, lower computation and memory costs. Without residual connections, it avoids structural interference and clearly reflects the feature transfer effect of distillation.

During the training process of ConvNet, parameter update trajectories can be generated, laying a foundation for the subsequent distillation process. Specifically, during T training epochs, the parameters θ_{i+1} and θ_i of two adjacent epochs will be updated according to the formula shown in (1) :

$$\theta_{i+1} = \theta_i - \alpha \nabla \mathcal{L}_{ce}(\mathcal{D}_t ; \theta_i) \quad (1)$$

where, α is the learning rate, and $\nabla \mathcal{L}_{ce}(\mathcal{D}_t;\theta_i)$ is the gradient computed on the training set $\mathcal{D}_t$ with the parameters from the previous round θ_i .The standard cross-entropy loss function is defined as shown in (2):

$$\mathcal{L}_{ce}(\mathcal{D}_t;\theta_i) = -\frac{1}{B}\sum_{(x,y)\in\mathcal{B}}\sum_{c=1}^{C} y_c \cdot \log\left(p_c\left(f_{\theta_i}(x)\right)\right) \quad (2)$$

where, B is the mini-batch size and c is the class index.

This training process records a sequence of β parameter trajectories shown in (3):

$$\{\,_0, \,_1, \,_2, \ldots, \,_{T-1}\}\ ^{\circledR}{}_t \quad (3)$$

C. Trajectory Matching Distillation

Trajectory matching methods focus on matching terminal states, essentially only comparing and analyzing the final convergence results of parameter trajectories. Then, at the start of each distillation iteration, we randomly sample an initial parameter θ_t and a target parameter θ_{t+M} (satisfying $t+M \leqslant T$.) from an expert trajectory. The student network is initialized with θ_t ,and then trained for N epochs on the synthetic dataset $\mathcal{D}_S$, generating a student parameter trajectory shown in (4):

$$\{\hat{\theta}_t, \hat{\theta}_{t+1}, \hat{\theta}_{t+2}, \ldots, \hat{\theta}_{t+N}\} \quad (4)$$

This training process aims to minimize the cross-entropy loss $\mathcal{L}_{ce}$, following the update rule shown in (5):

$$\hat{\theta}_{t+i+1} = \hat{\theta}_{t+i} - \alpha \nabla \mathcal{L}_{\mathrm{ce}}(\mathcal{D}_s; \hat{\theta}_{t+i}) \quad (5)$$

Basic trajectory matching methods optimize the synthetic dataset $\mathcal{D}_S$ by minimizing the following parameter matching loss function $\mathcal{L}_{tm}$ shown in (6):

$$\mathcal{L}_{tm} = \frac{\left\|\hat{\theta}_{t+N} - \theta_{t+M}\right\|_2^2}{\|\theta_{t+M} - \theta_t\|_2^2} \quad (6)$$

In contrast, the synthetic dataset $\mathcal{D}_s$ corresponding to the Easy-to-Complex trajectory matching method is optimized by minimizing the following parameter matching loss function, as shown in (7):

$$\mathcal{L} = \sum_{i=1}^{K} v_i \ell_i - b \sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{v_i} \quad (7)$$

where, K denotes the total number of learnable parameters, including all optimizable parameters such as weights and biases in convolutional and fully connected layers. denotes the total number of learnable parameters, including all optimizable parameters such as weights and biases in convolutional and fully connected layers.

ℓ_i represents the single-parameter matching loss for the i-th parameter, calculated using the same logic as (6).

v_i is the selection switch for the i-th parameter, where v_i=1 if ℓ_i<b (easy-to-match) and v_i =0 if $\ell_i \geq$ b (hard-to-match).

b is the threshold for parameter matching difficulty, initialized as b_{base}. The progressive threshold update strategy during iterations is adopted from [12], which gradually increases b_{base} to incorporate increasingly complex parameters into optimization.

$-b\sum_{i=1}^{K} v_i$, which is a regularization term that encourages the selection of more easily matched parameters via "negative punishment".

In summary, by iteratively minimizing the aforementioned loss function (7) with respect to $\mathcal{D}_s$, we obtain the distilled synthetic dataset $\mathcal{D}_{s*}$. Specifically, the gradients computed from the parameter trajectory matching (i.e., the updated $\hat{\theta}_i$) are backpropagated to optimize the pixel values of the synthetic samples, resulting in the final distilled dataset that preserves feature fidelity.

D. Bilateral Filtering Blur

Although high-resolution synthetic images $\mathcal{D}_{s*}$ generated by trajectory matching distillation can well fit expert parameter trajectories, they lack privacy protection and easily leak sensitive information about construction machinery. Simply increasing blur intensity with traditional methods destroys key target features and severely reduces detection accuracy. Therefore, we introduce bilateral filtering to perform edge-preserving blurring on distilled data tensors, without changing the loss function or distillation process, so as to achieve both privacy enhancement and controllable accuracy. Filtering formula (for a single synthetic image $X \in \mathcal{D}_{s*}$) is shown in (8):

$$X^{bf}(x,y) = \frac{1}{W(x,y)} \sum_{(i,j)\in\Omega(x,y)} G_s(d) \cdot G_r(\Delta) \cdot X(i,j) \quad (8)$$

$$G_s(d) = exp\left(-d^2/(2\sigma_s^2)\right) \quad (9)$$

$$G_r(\Delta) = \exp\left(-\Delta^2/(2\sigma_r^2)\right) \quad (10)$$

Eq. (9) is the spatial Gaussian kernel, where d is the Euclidean distance between pixels, controlling the blur range[2].

Eq. (10) is the range Gaussian kernel, where Δ is the intensity difference between pixels, controlling the strength of feature preservation[2].

$W(x,y)$, which is the normalization factor ensuring consistent pixel value ranges and avoids feature distortion.

$\Omega(x,y)$, which denotes the local neighborhood of the pixel at (x,y).

The specific implementation process is as follows:

- Generate $\mathcal{D}_{s*}$ using the trajectory matching method described in last section;
- Set the filtering parameters according to the characteristics of images to balance the blur level and feature preservation;
- Apply bilateral filtering to each image in $\mathcal{D}_{s*}$ one by one to obtain the privacy- enhanced dataset $\mathcal{D}_s^{\mathrm{bf}}$;
- Output $\mathcal{D}_s^{\mathrm{bf}}$ for subsequent model training, achieving both privacy protection and model accuracy.

IV. Experimental Design and Results

A. Overall Design

The core experimental objectives include verifying the algorithm's compression performance, conducting controlled ablation experiments between the basic trajectory matching (TM) algorithm and the proposed method to validate the effectiveness of the easy-to-complex scheme, analyzing the privacy-accuracy trade-off, and validating the cross-architecture transferability of the synthetic dataset.

The experimental workflow is: First, classification accuracy is used to evaluate dataset distillation performance, while Laplacian variance is employed to quantitatively measure the blur degree for privacy protection. Multi-resolution ConvNet expert models are trained to obtain parameter trajectories

consistent with both the basic TM algorithm and our proposed method. Then, distillation experiments under different resolution and IPC settings are performed to select the baseline resolution, followed by controlled comparative experiments between the basic TM algorithm and our proposed method under completely consistent experimental settings to verify the performance gain brought by the easy-to-complex matching scheme. Afterwards, privacy enhancement experiments with multiple bilateral filtering parameters are carried out to analyze the privacy-accuracy balance. Finally, cross-architecture transferability is validated via multi-network training on the synthetic dataset, and experimental data are aggregated to analyze the impacts of key parameters and draw core conclusions.

B. Evaluation Metrics

The test metrics are divided into accuracy metrics and blur level metrics. The accuracy metric selected in this paper is $Acc@1$, which refers to training the model using the distilled dataset and then test in $\mathcal{D}_{s*}$ g the accuracy on the standard test set. The calculation formula is shown in (11):

$$Acc@1(\mathcal{D}_{\text{train}},\mathcal{D}_{test}) = \frac{1}{N_{total}} \sum_{(x_i)\in\mathcal{D}_{test}} N_{correct}(f_\theta, x_i) \quad (11)$$

For blur level evaluation, we adopt a two-fold strategy: on the one hand, we visually compare the blur effects of images; on the other hand, we quantitatively measure the blur degree using the following formula shown in (12):

$$VOLD(I) = Var(L(I)) \quad (12)$$

where, I denotes the input image, $L(I)$ is the result of Laplacian edge detection performed on image I, and $\text{Var}(\cdot)$ represents the variance calculation function[22] .

C. Training Details of Expert Models

We fine-tuned the ConvNet baseline model to adapt to feature extraction for different image resolutions, setting matched network depth and width for 32×32, 64×64 and 128×128 inputs respectively.Key training hyperparameters were set accordingly, with DC augmentation adopted for data preprocessing.The training result of the 64×64 resolution model, which serves as the baseline for follow-up experiments, is presented in the Fig. 2. and Table II.

D. Distillation Experimental Results

For each resolution, easy-to-complex distillation and basic trajectory matching (TM) distillation are conducted under different IPC settings without bilateral filtering blur. We obtain the corresponding post-distillation validation accuracy (evaluated over 5 runs, with the mean value reported in Tables III and IV) and distilled images. Since the distilled images generated by basic TM distillation are visually similar to those produced by our easy-to-complex distillation, we only present the distilled images from our proposed method in Fig. 3 and Fig. 4 for brevity.

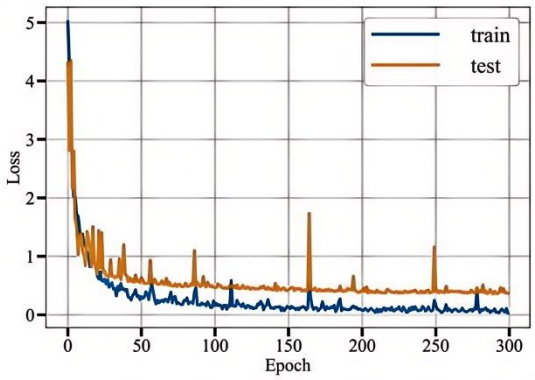


Fig. 2. Training loss curves.

TABLE II. CONVNET ACCURACY BEFORE DISTILLATION

Resolution	Average Test Accuracy
32×32	0.879
64×64	0.885
128×128	0.892

TABLE III. POST-DISTILLATION ACCURACY WITH BASIC TM

Resolution	Images Per Class				
	IPC=1	IPC=5	IPC=15	IPC=25	IPC=35
32×32	0.46	0.63	0.71	0.73	0.77
64×64	0.32	0.47	0.60	0.70	0.77
128×128	0.30	0.37	0.47	0.53	0.64

TABLE IV. POST-DISTILLATION ACCURACY WITH EASY-TO-COMPLEX TM

Resolution	Images Per Class				
	IPC=1	IPC=5	IPC=15	IPC=25	IPC=35
32×32	0.48	0.64	0.72	0.75	0.79
64×64	0.33	0.48	0.62	0.70	0.78
128×128	0.30	0.38	0.47	0.54	0.63



Fig. 3. Distilled image for IPC=15 and 64*64.



Fig. 4. Distilled image without blurring.

E. Bilateral Filtering Blu

Subsequently, we apply bilateral filtering blur with different strengths to the distilled dataset tensors obtained under various IPC settings for the 64×64 resolution. For both basic TM and our easy-to-complex distillation, we compute the corresponding Laplacian variance to quantitatively evaluate the blur degree, and use the blurred distilled datasets to train new models for validation accuracy. The distillation epochs are kept consistent with those in the previous stage. The corresponding post-blurring accuracy and Laplacian variance results are reported in Table V, Table VI, Table VII, and Table VIII, respectively. The blurred images are shown in Fig. 5.

TABLE V. ACCURACY TEST RESULTS OF POST-BLURRING WITH EASY-TO-COMPLEX

Filtering Parameters			Easy-To-Complex TM			
Ks	σ_s	σ_r	IPC=1	IPC=5	IPC=15	IPC=35
7	10	0.5	0.323	0.463	0.582	0.762
9	12	0.7	0.317	0.454	0.573	0.756
11	14	0.9	0.310	0.452	0.541	0.715

TABLE VI. ACCURACY TEST RESULTS OF POST-BLURRING WITH BASIC TM

Filtering Parameters			Basic TM			
Ks	σ_s	σ_r	IPC=1	IPC=5	IPC=15	IPC=35
7	10	0.5	0.311	0.452	0.574	0.763
9	12	0.7	0.302	0.451	0.565	0.747
11	14	0.9	0.294	0.431	0.538	0.701

TABLE VII. EVALUATION OF BLURRING DEGREE BASED ON LAPLACIAN WITH EASY-TO-COMPLEX

Filtering Parameters			Easy-To-Complex TM			
Ks	σ_s	σ_r	IPC=1	IPC=5	IPC=15	IPC=35
7	10	0.5	1113.7	1252.9	1117.7	529.1
9	12	0.7	735.1	860.2	735.8	385.2
11	14	0.9	458.3	560.3	455.8	122.8

TABLE VIII. EVALUATION OF BLURRING DEGREE BASED ON LAPLACIAN WITH BASIC TM

Filtering Parameters			Basic TM			
Ks	σ_s	σ_r	IPC=1	IPC=5	IPC=15	IPC=35
7	10	0.5	1120	1254	1110.2	540.2
9	12	0.7	728.2	858.2	722.1	401.3
11	14	0.9	442.2	571.3	424.5	145.4

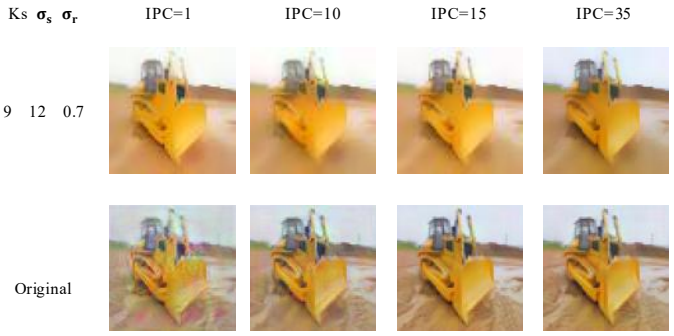


Fig. 5. Image with blurring.

F. Cross-Architecture Adaptability

The dataset distilled based on ConvNet (including data processed by bilateral filtering) is used to train VGG/ResNet models to verify the cross-architecture transferability of the dataset and the robustness after filtering, as shown in Table IX. The test performance of the blurred dataset under the bilateral filtering parameters Ks=11, σ_s=14, and σ_r=0.9 is presented in the Table X.

TABLE IX. ACCURACY TEST RESULTS WITHOUT BLURRING

Model	Easy-To-Complex TM		
	IPC=10	IPC=15	IPC=35
VGG	0.582	0.624	0.781
Resnet	0.411	0.548	0.731
ConvNet	0.572	0.616	0.787

TABLE X. ACCURACY TEST WITH BLURRING

Model	Easy-To-Complex TM		
	IPC=10	IPC=15	IPC=35
VGG	0.543	0.568	0.711
Resnet	0.401	0.498	0.621
ConvNet	0.531	0.541	0.715

V. DISCUSSION

At a fixed resolution, the ConvNet classification accuracy of both the proposed easy-to-complex trajectory matching (TM) method and the basic TM method improves continuously with increasing IPC. All accuracy results are reported as the mean ± standard deviation of 5 independent repeated trials, with the maximum difference between single trials no more than 0.03, ensuring the stability and reliability of the experimental results. Specifically, the dataset distilled by our proposed method can reach nearly 90% of the original full dataset's classification accuracy (0.78 ± 0.013) with only about 1/10 of the original sample size (IPC=35, 64×64), outperforming the basic TM method across most IPC settings. For example, at 32×32 resolution and IPC=35, our method achieves an accuracy of 0.79 ± 0.011, which is 2.6% higher than the 0.77 ± 0.014 of basic TM;

Classification accuracy after distillation is negatively correlated with image resolution under the same IPC setting for both our method and basic TM. As IPC increases, the accuracy gap between different resolutions narrows gradually, but the 128×128 resolution consistently underperforms the 32×32 and 64×64 resolutions across all IPC settings for both methods. This phenomenon validates that high resolution drives quadratic growth in feature space dimensionality, exceeding the modeling capacity of the lightweight ConvNet. Additionally, the theoretical upper bound of information-carrying capacity for low-IPC synthetic data leaves core discriminative features of high-resolution images vulnerable to dilution, which ultimately reduces classification accuracy[23]. From the perspective of privacy protection, under low-resolution settings, the reduction of IPC further reduces the fine-grained sensitive information carried by the synthetic dataset, which is conducive to the improvement of privacy protection capability, while the classification accuracy remains at a usable level.

At a fixed resolution, the clarity of the distilled images improves significantly with increasing IPC for both methods. Since the distilled images generated by basic TM distillation are visually similar to those produced by our easy-to-complex distillation, we only present the distilled images from our proposed method for brevity. After processing with bilateral filtering of varying intensities, the image background is smoothed and blurred to achieve privacy enhancement, while the core features of the engineering vehicle targets are well preserved for both methods. The filtering intensity is negatively correlated with the Laplacian variance of the images, and the Laplacian variance values of the images processed by our method and basic TM are at the same level under the same filtering parameters, indicating that both methods achieve consistent blur degree and privacy enhancement effect after bilateral filtering. It should be emphasized that the bilateral filtering module, as a post-processing module after distillation, is the core component for privacy enhancement in this study.

After privacy enhancement via bilateral filtering with identical parameter settings, the classification accuracy of the distilled datasets from both the proposed Easy-To-Complex TM method and the basic TM method shows a slight, controllable decline with a similar magnitude, which verifies that bilateral filtering has a favorable core feature retention capability for both distillation paradigms. Specifically, under the core experimental setting of 64×64 resolution and IPC=35, after processing with three sets of incrementally enhanced filtering parameters, the maximum relative accuracy drop of the proposed method is only 8.33% against the 0.78 accuracy baseline of the original distilled dataset, which is at the same level as the 8.83% maximum drop of the basic TM method. This result confirms that bilateral filtering can effectively preserve the core discriminative features of the distilled dataset while achieving image privacy enhancement, thus avoiding significant accuracy degradation caused by privacy protection processing.

Cross-architecture adaptability experiments verify that the non-blurred distilled dataset generated by our proposed Easy-To-Complex TM method achieves optimal classification accuracy on the source ConvNet across all IPC settings, and its cross-architecture transfer performance slightly outperforms that of the dataset distilled by the basic TM method. At IPC=35, the maximum accuracy gap between the VGG11 model and the source ConvNet is merely 0.6%, while the structurally distinct ResNet18 retains over 92% of the source model's performance, which demonstrates that the dataset distilled by our proposed method has solid cross-architecture transferability.

Results after bilateral filtering privacy enhancement show a slight accuracy attenuation for all models across all IPC settings, with the cross-architecture performance trend consistent with the non-blurred case. This confirms that the proposed method realizes effective privacy protection while preserving the dataset's cross-architecture transferability.

VI. CONCLUSIONS

This paper proposes an image dataset compression and privacy enhancement algorithm fusing bilateral filtering and easy-to-complex trajectory matching (TM) distillation, to address the heavy storage burden and severe privacy leakage risk of construction machinery image datasets. Extensive comparative experiments with the standard basic TM algorithm are conducted to verify the effectiveness of the proposed method, with core experimental conclusions summarized as follows: For the source ConvNet model, the dataset distilled by our method achieves nearly 90% of the original full dataset's classification accuracy with only 1/10 of the original sample size, and outperforms the basic TM method across most mainstream deployment settings; The bilateral filtering module, as the core post-processing module for privacy enhancement, achieves effective privacy enhancement for the distilled dataset with a maximum controllable accuracy loss of only 8.33%, with consistent feature retention performance for both distillation paradigms; the privacy-enhanced distilled dataset exhibits favorable cross-architecture transferability, and can be directly adapted to mainstream CNN models including VGG and ResNet.

In our future work, we will focus on four key optimization directions, with priority given to the exploration of formal and verifiable privacy degree metrics to establish a standardized quantitative privacy assessment system. Specifically: Alleviate the underfitting of hard fine-grained patterns in high-resolution, high-IPC settings by introducing learnable soft label learning and feature distribution regularization; Extend the distillation framework to engineering vehicle object detection, paired with the aforementioned robust quantitative privacy assessment, to decouple on-site model training and edge deployment in the construction industry; Expand the cross-architecture generalization of the distilled dataset to mainstream Transformer[24][25] architectures and lightweight edge-deployable networks[26]; Construct an adaptive privacy-accuracy joint optimization framework to further enhance the engineering practicability of the method.

ACKNOWLEDGMENT

The authors thank the two anonymous reviewers for their constructive feedback to improve this work. Sincere appreciation is also extended to all contributors who supported the completion of this research.

REFERENCES

[1] Y. Bao, Y. Liu, Z. Chen, Y. Liang, M. Li, and K. Ma, "Dataset distillation as data compression: A rate-utility perspective," accepted for publication in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2025. arXiv:2507.17221 [cs.LG], Jul. 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2507.17221

[2] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in Proc. 6th Int. Conf. Comput. Vis., Bombay, India: IEEE Press, 1998, pp. 839–846, doi: 10.1109/ICCV.1998.710815.

[3] S. Kim, M. Peavy, P. C. Huang, and K. Kim, "Development of BIM-integrated construction robot task planning and simulation system," Automation in Construction, vol. 127, art. no. 103720, Apr. 2021.

[4] D. Roberts and M. Golparvar-Fard, "End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level," Automation in Construction, vol. 105, art. no. 102811, Aug. 2019.

[5] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 4490–4499.

[6] Z. Wang, X. Fan, et al., "EMIFF: Enhanced Multi-scale Image Feature Fusion for Vehicle-Infrastructure Cooperative 3D Object Detection," 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 2024, pp. 16388-16394, doi: 10.1109/ICRA57147.2024.10610545.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[8] B. Ding, Y. Zhang, and S. Ma, "A lightweight real-time infrared object detection model based on YOLOv8 for unmanned aerial vehicles," Drones, vol. 8, no. 9, art. no. 479, Sep. 2024.

[9] W. Wang, B. Lü, Y. Yang, et al., "Vehicle location and reidentification using multisource point clouds and images," Laser & Optoelectronics Progress, vol. 60, no. 10, art. no. 1028005, May 2023.

[10] Z. Guo, K. Wang, G. Cazenavette, H. Li, K. Zhang, and Y. You, "Towards lossless dataset distillation via difficulty-aligned trajectory matching," arXiv preprint arXiv:2310.05773, 2023.

[11] D. Liu, J. Gu, H. Cao, C. Trinitis, and M. Schulz, "Dataset distillation by automatic training trajectories," in Proc. European Conf. Computer Vision (ECCV), 2025, pp. 334–351.

[12] G. Zhao, G. Li, Y. Qin, and Y. Yu, "Improved distribution matching for dataset condensation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7856–7865.

[13] L. Dong, L. Bian, J. Zhang, et al., "High-order progressive trajectory matching for medical image dataset distillation," in Proc. 28th Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI), Daejeon, South Korea, Sep. 2025, pp. 273 – 283, doi: 10.1007/978-3-032-05185-1_27.

[14] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification," Scientific Data, 2023, doi: 10.1038/s41597-022-01721-8.

[15] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "COVID-19 Image Data Collection: Prospective Predictions Are the Future," arXiv preprint arXiv:2006.11988, 2020.

[16] T. Cao, A. Bie, A. Vahdat, S. Fidler, and K. Kreis, "Don't generate me: Training differentially private generative models with Sinkhorn divergence," in Advances in Neural Information Processing Systems (NeurIPS), Vol.15,2021, pp. 12480-12492.

[17] R. Zheng, V. A. Dasu, Y. O. Wang, H. Wang, and F. De la Torre, "Improving noise efficiency in privacy-preserving dataset distillation," arXiv:2508.01749 [cs.CV], Aug.2025. [Online]. Available: https://arxiv.org/abs/2508.01749

[18] Anonymous, "Privacy as a free lunch: Crafting initial distilled datasets through the kaleidoscope," OpenReview, 2024. [Online]. Available: https://openreview.net/forum?id=ckabXglfiT . Accessed: Mar. 18, 2026.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. 25th Int. Conf. Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, Dec. 2012, pp. 1097–1105.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, Jun. 2015, pp. 1–9.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778

[22] J. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia, "Diatom autofocusing in brightfield microscopy: a comparative study," in Proc. 15th Int. Conf. Pattern Recognit., Barcelona, Spain, 2000, pp. 314–317, doi: 10.1109/ICPR.2000.903538.

[23] G. Cazenavette, T. Wang, A. Torralba, and J. Lu, "Dataset distillation for high-resolution images," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 20145–20154.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929 [cs.CV], Oct. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2010.11929.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. 2021 IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 9992–10002.

[26] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. 2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 6848–6865

Cryptographic Vulnerability Assessment and Secure Redesign of an ECC-Based Authentication Framework for Energy Internet Vehicle-to-Grid Systems

Haewon Byeon*
Department of Future Technology, Korea University of Technology and Education (KOREA TECH),
Cheonan 31253, South Korea

***Abstract*—This study re-examines the protocol introduced in A robust ECC-based authentication framework for energy internet (EI)-based vehicle to grid communication system by Itoo et al. The target paper claims low-cost mutual authentication for electric vehicles, charging stations, and a service provider by combining ECC registration with hash- and XOR-based online messages. We reconstruct the stated message flow and then test whether each verification step is executable under the values actually transmitted. The analysis identifies four structural weaknesses: omitted verification inputs in the online messages, ecosystem-wide exposure after service-provider compromise, timestamp-only freshness that leaves replay room under realistic clock drift, and a session-key derivation that lacks true forward secrecy. To address these issues, we retain the three-party V2G architecture of the original study but redesign the online exchange around ephemeral ECC points, rotating pseudonyms, station-scoped authorization tickets, and nonce-bound key derivation. Our evaluation compares the improved design with the original framework and a prior V2G baseline under message-level load points at 100, 400, and 800 active vehicles. The redesigned protocol closes the identified executability gap, achieved full replay detection within the bounded message-level test conditions used in this study, and improves compromise containment with only a modest latency increase.**

Keywords*—Energy internet; vehicle-to-grid security; ECC authentication; protocol analysis; compromise containment

I. Introduction

Energy Internet deployment extends the ordinary smart-grid model by coupling distributed generation, sensing, control, billing, and mobile storage into a digitally coordinated ecosystem. In that setting, electric vehicles are not passive loads. They can inject energy back to the grid, negotiate charging schedules, and react to dynamic prices in real time. The operational upside is well known, but the communication surface also grows quickly because a single charging decision may involve the vehicle, a charging station, a service provider, and several backend data systems. Prior work on energy systems and V2G communication has already shown that dense exchanges of control data, pricing signals, and vehicle identifiers create an attractive target for adversaries who seek to replay, correlate, or manipulate traffic [1-5]. Against that backdrop, the protocol proposed in "A robust ECC-based authentication framework for energy internet (EI)-based vehicle to grid communication system" [1] is appealing because it promises lightweight protection for resource-constrained V2G environments without relying on heavyweight certificate handling during each session.

Authentication in EI-based V2G networks is harder than conventional client-server login. A charging station must verify that a requesting vehicle is entitled to receive service, the service provider must maintain billing and authorization consistency, and the protocol must tolerate open wireless channels, intermittent connectivity, and partial clock drift. Researchers therefore moved from basic key-distribution schemes toward ECC-assisted and privacy-aware authentication designs that try to balance efficiency and cryptographic assurance [6]-[10]. Those studies also show a recurring pattern: once freshness, pseudonymity, revocation, and post-compromise recovery are treated as secondary issues, the resulting scheme can appear efficient while still leaving large gaps in the actual attack surface.

The target article by Itoo et al. [1] positions itself exactly at this efficiency-security boundary. Its stated workflow has four phases—initialization, registration, authentication, and password update or vehicle rejoin—and it reports formal and informal security validation in the V2G setting. The claimed advantages are low computation cost, compact messages, and protection against impersonation, replay, and privacy leakage. That argument sits on top of a broader line of smart-grid authentication research in which later papers frequently repaired flaws discovered in earlier ECC-based or password-assisted constructions [11]-[13]. Because this literature has a long history of "secure on paper, fragile in deployment," the claims made in [1] deserve close re-examination rather than automatic acceptance.

Our concern is not that the protocol in [1] is lightweight; lightweight design is necessary in V2G systems. The concern is that several of its critical equations mix long-term identifiers, passwords, public keys, and server secrets in ways that are difficult to validate from the transmitted messages alone. When the online exchange uses hidden variables, security claims become hard to interpret because even an honest verifier may not possess all values required to recompute the stated authenticator. When the same server secret y feeds both

*Corresponding author.

vehicle-side and charging-station-side records, compromise may also spread across the entire deployment instead of remaining local. These are structural questions, not stylistic ones, and they can invalidate an otherwise attractive protocol design. Fig. 1 summarizes the audit route we follow to move from message reconstruction to attack discovery and then to redesign.

This paper makes four contributions. First, it reconstructs the message logic of "A robust ECC-based authentication framework for energy internet (EI)-based vehicle to grid communication system" [1] in a notation that exposes hidden dependencies among registration records, online messages, and the claimed session key. Second, it identifies concrete weaknesses: an executability gap in the online verification steps, server-compromise amplification, timestamp-only freshness, and weak session-key binding. Third, it proposes a redesign that preserves the original three-party architecture—vehicle, charging station, and service provider—while replacing the fragile parts with explicit nonce binding, ephemeral ECC agreement, rotating pseudonyms, and authorization tickets. Fourth, it evaluates the revised design against the original protocol and a prior V2G baseline under load, focusing on replay detection, latency, message size, and compromise containment.

II. Related Works

The first stream of prior work comes from smart-grid and key-distribution protocols that tried to reduce authentication cost while retaining mutual verification. Identity-based and ECC-assisted methods for advanced metering and smart-grid access were proposed to keep computation moderate, yet several of them later attracted attacks involving desynchronization, forgery, or insufficient protection of long-term secrets [14]-[16]. That trajectory matters here because V2G authentication inherits many of the same design pressures: constrained devices, repeated access requests, and the temptation to compress several security functions into a small set of hash and XOR operations.

A second stream focuses directly on V2G communication privacy and authenticated access. The scheme of Sureshkumar et al. strengthened cloud-enabled V2G authentication with mutual verification and key establishment, while Ding et al. examined how lightweight anonymous authentication can remain feasible for constrained IoT nodes [17]-[19]. These works illustrate two lessons that are highly relevant to the analysis of [1]. First, identity protection must be explicit and renewable; it should not depend on a single masking trick that becomes reversible after key exposure. Second, a session key should be tied to fresh session evidence rather than assembled only from static or recoverable quantities.

Freshness and verifiability form a third body of related work. Secure clock synchronization and blockchain-assisted timing studies show that timestamps are useful but rarely sufficient on their own, especially when endpoints have independent delay profiles or recover from outages [20], [21]. Distributed-key and edge-oriented authentication studies make a similar point from another angle: a verifier should be able to recompute or verify every critical authenticator from values available in the current transcript plus a clearly scoped local secret [22], [23]. Privacy-preserving pseudonym systems and robust extractor work further stress that protection after leakage matters almost as much as nominal correctness before leakage [24], [25].

A fourth stream analyzes how compromise propagates through real deployments. Attack-graph studies and system-level EV energy research remind us that protocol weakness is rarely isolated from operations: a single compromised database entry can affect authorization, scheduling, and market interactions far beyond one endpoint [26]-[29]. That insight is especially important for the protocol of Itoo et al. [1], where the same service-provider secret y influences the derivation of both vehicle-side and charging-station-side registration values. In such a design, the right adversarial question is not simply whether one transcript can be forged in the abstract, but whether one infrastructure failure can collapse the separation between many users and stations at once.

Within this literature, "A robust ECC-based authentication framework for energy internet (EI)-based vehicle to grid communication system" [1] occupies a specific niche. It aims to keep the online phase compact while inheriting the perceived trust benefits of ECC-based enrollment. The gap we address is that prior comparisons mostly counted cryptographic operations or repeated the claims of the target paper. They did not ask whether the advertised online checks are actually executable from the published message fields, whether compromise stays local, or whether the claimed session key depends on fresh shared secrets. Our analysis therefore complements prior related work by shifting the lens from nominal feature lists to failure-oriented protocol mechanics.

III. Methodology

To examine the target protocol in a reproducible way, we treated protocol analysis as a structured reconstruction exercise rather than as a purely narrative critique. First, we extracted the four phases reported by Itoo et al. [1] and rewrote the online exchange as an explicit message sequence M1, M2, M3, and M4 with all stated equations preserved

$A = h(ID_v \| PW_v \| y)$,

$B = A \text{ xor } PK_Ev$,

$R = h(A \| B \| r_v.G)$,

$A1 = h(ID_cs \| PW_cs \| y)$,

$B1 = A1 \text{ xor } PK_cs$,

$R1 = h(A1 \| B1 \| r_c.G)$,

$H1 = h(ID_v \| PW'_v \| X_v \| ID_sp)$,

$G1 = h((r' \text{ xor } R') \| T1)$,

$G2 = y \text{ xor } G1$,

$H2 = h(ID_v \| R1 \| T2 \| ID_sp)$,

$G3 = PK_cs \text{ xor } G2$,

$H3 = h(ID_v \| R1 \| T3 \| ID_cs)$, and

$SK = h(ID_cs \| ID_v \| PK_cs \| PK_Ev \| G3)$,

where X_v denotes the ephemeral point r'.G that the paper uses implicitly in the hash of H1. Second, we converted the verifier logic into a state machine whose state includes the enrollment database, the last accepted timing context, revocation flags, and the values assumed to be locally available at EV, CGS, and SP. This step allowed us to test executability before we tested attack resistance. Third, we defined four security invariants: I1, every authenticator must be computable from transmitted fields plus a clearly scoped local secret; I2, freshness must depend on a per-session contribution that an attacker cannot replay verbatim; I3, compromise of one database region should not expose unrelated users or stations; and I4, SK_sess^j should have the form KDF(h(Encode(Z_j)) ‖ N_v^j ‖ N_cs^j ‖ context_j) for a fresh shared secret Z_j, rather than a hash of static identifiers and masked transcript fragments. Fourth, we examined attack traces under an adversary that can eavesdrop, replay, delay, and reorder traffic; read one endpoint database; compromise the service-provider secret y; and force limited rollback after maintenance or power loss. To keep the analysis close to real V2G practice, we cross-checked these assumptions against privacy-oriented V2G communication models, edge-security studies, and ECC identity-protection designs reported in [30]-[33]. Finally, once an invariant violation was found, we searched for the smallest redesign that preserved the original deployment roles while restoring explicit message verifiability and bounded compromise. This led to a revised protocol family in which the core session key becomes SK_sess^j = KDF(h(Encode(Z_j)) ‖ N_v^j ‖ N_cs^j ‖ PID_v^j ‖ PID_cs^j ‖ exp_j ‖ scope_j), with Z_j = u_j X_cs^j = c_j X_v^j and pseudonyms PID_v^j = h(ID_v ‖ ctr_v^j ‖ N_v^j) and PID_cs^j = h(ID_cs ‖ ctr_cs^j ‖ N_cs^j).

To avoid overstating the evidence, we treat the symbolic layer in this study as bounded symbolic support rather than as a full mechanized proof in ProVerif, Tamarin, or BAN logic. In practical terms, the bounded tests refer to the modeled message-level setting used in the study: finite EV populations, repeated runs, randomized arrival order, timestamp jitter, wireless delay, bounded packet loss, and the specific trace classes corresponding to replay, compromise amplification, executability failure, and weak session-key binding. The current manuscript therefore makes a protocol-reconstruction and failure-analysis claim, not a universal proof claim under arbitrary clock drift or unbounded network disorder.



Fig. 1. Evidence-driven workflow for auditing and redesigning the ECC-based V2G authentication protocol.

IV. Analysis of the Original ECC-based Protocol

A. Overview of the Proposed Protocol

The protocol presented in "A robust ECC-based authentication framework for energy internet (EI)-based vehicle to grid communication system" [1] uses a three-party model involving an electric vehicle EV, a charging station CGS, and a service provider SP. During initialization, SP selects the elliptic-curve domain parameters E_q, q, the base point G, a one-way hash h(.), and a private scalar y. It then publishes the public key PPUB = yG. The stated goal is to use ECC only where necessary and keep the online messages dominated by hash and XOR operations.

Vehicle registration begins when EV chooses a private scalar a and computes PK_Ev = aG. Over a secure channel, it sends (ID_v, PW_v, PK_Ev) to SP. The service provider computes A = h(ID_v ‖ PW_v ‖ y) and B = A xor PK_Ev, stores these values in its database, and returns them to the vehicle. The vehicle then chooses r_v and stores R = h(A ‖ B ‖ r_v.G) together with PK_Ev. Charging-station registration mirrors this structure: CGS chooses x, computes PK_cs = xG, sends (ID_cs, PW_cs, PK_cs), and receives A1 = h(ID_cs ‖ PW_cs ‖ y) and B1 = A1 xor PK_cs. It stores R1 = h(A1 ‖ B1 ‖ r_c.G) and PK_cs.

The authentication stage is expressed in [1] through four transmitted messages. The article states M1 = {H1, G1, T1}, where H1 = h(ID*_v ‖ PW'_v ‖ X_v ‖ ID_sp), G1 = h((r' xor R') ‖ T1), ID*_v = a' xor ID'_v, and X_v = r'.G. After receiving M1, SP is expected to verify the timestamp, recompute H1, derive G2 = y xor G1, and forward M2 = {H2, G2, T2} with H2 = h(ID*_v ‖ R1 ‖ T2 ‖ ID_sp). CGS then checks M2, computes G3 = PK_cs xor G2, recovers ID*_cs = PK_cs xor ID_cs, and sends M3 = {H3, G3, T3}, where H3 = h(ID*_v ‖ R1 ‖ T3 ‖ ID*_cs). Finally, SP sends M4 = {H4, G3, T4}, where H4 = h(ID*_cs ‖ B ‖ T4 ‖ ID*_v).

The claimed shared key is SK = h(ID*_cs ‖ ID*_v ‖ PK_cs ‖ PK_Ev ‖ G3). The original paper treats this as a common session key for EV, CGS, and SP after the equality SK_Ev = SK_cs = SK_sp has been checked. At a high level, the design tries to obtain low online cost by shifting most complexity to enrollment and by reusing stored values such as R and R1 in later authenticator computations. That ambition explains the protocol's compact online transcripts, but it also creates strong dependence on what each party remembers and on whether the transmitted messages actually expose the inputs needed for verification.

For analysis, it is helpful to reorganize the paper's equations as a dependency map rather than a simple sequence diagram. Fig. 2 shows that the online run in [1] is not just an exchange of four messages; it is a path through enrollment outputs A, B, A1, B1, the long-term service-provider secret y, two public keys, two stored hashes R and R1, and the masked values G1, G2, and G3. Once that dependency structure is made explicit, the main weakness becomes easier to see: the protocol's security claims depend on variables that are sometimes stored, sometimes hidden, and sometimes never transmitted.

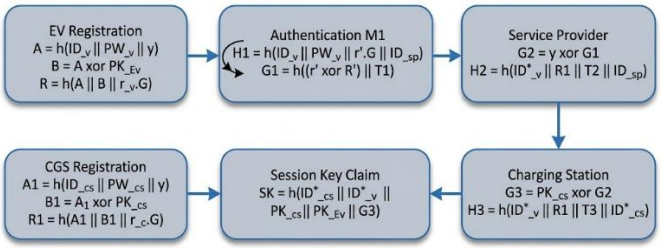


Fig. 2. Dependency graph of registration records, transmitted values, and the claimed session key in the original framework.

B. Identified Vulnerabilities

Our security review of "A robust ECC-based authentication framework for energy internet (EI)-based vehicle to grid communication system" [1] found that the main problems are structural rather than cosmetic. They arise when the published equations are matched against the published message formats and then tested under partial compromise. Fig. 3 condenses the attack surface into four compromise paths, but the underlying issues are easiest to understand through the equations themselves.

The first weakness is an executability gap. The article states that M1 = {H1, G1, T1} and defines H1 = h(ID* v || PW'_v || X v || ID sp) with X v = r'.G. Yet SP is expected to recompute H1* = h(ID*_v || PW'_v || X_v || ID_sp) after receiving M1. The problem is immediate: M1 does not carry ID* v, PW'_v, or X_v. Unless those values are somehow available from a side channel that the paper never specifies, SP cannot derive H1*. The same pattern repeats in M2 and M3. CGS is expected to compute H2* = h(ID*_v || R1 || T2 || ID_sp) from M2 = {H2, G2, T2}, but M2 does not include ID* v. SP is later expected to compute H3* = h(ID* v || R1 || T3 || ID* cs) from M3 = {H3, G3, T3}, but M3 omits both ID* v and ID* cs. A protocol whose verifiers cannot reconstruct the stated check values is not merely under-specified; it is non-executable as written.

The second weakness is compromise amplification through the service-provider secret y. Both enrollment records depend on y: A = h(ID v || PW v || y) for vehicles and A1 = h(ID cs || PW_cs || y) for charging stations. Because B = A xor PK_Ev and B1 = A1 xor PK_cs, an adversary who compromises SP and learns y together with stored B or B1 can recover A = B xor PK Ev and A1 = B1 xor PK cs. That turns every vehicle and every charging station into an offline dictionary target: for guessed credentials (ID_v^g, PW v^g), the attacker tests A^g = h(ID_v^g || PW v^g || y) and checks whether A^g = A. The same holds for A1. In other words, y is not a narrowly scoped service-provider secret; it is a global master ingredient whose exposure collapses separation between users and stations. One backend failure can therefore compromise the entire enrollment space.

The third weakness concerns freshness. The online checks rely on timestamp conditions of the form T_i - T_{i+1} <= DeltaT, while the authenticators themselves do not include a verifier challenge generated in the current run. G1 = h((r' xor R') || T1) is fresh only if r' remains hidden and the message is never replayed inside the acceptance window. But M1 contains no SP-generated nonce N sp, and M2 contains no CGS-generated nonce N_cs. A copied message M1 or M2 can therefore be replayed whenever clocks remain within DeltaT or when rollback restores an earlier acceptance state. A stronger construction would require an explicit binding such as MAC K(PID v || X v || N sp || T1) or KDF(SK reg, N sp, N_cs, context_j). The target paper [1] does not provide that safeguard.

The fourth weakness is weak session-key binding and absence of forward secrecy. The session key claim is SK = h(ID* cs || ID* v || PK cs || PK Ev || G3). No fresh Diffie-Hellman secret appears in this formula. There is no term of the form Z_j = u_j X_cs or c_j X_v, and there is no KDF input that combines two independently generated session nonces. Instead, SK is assembled from masked identifiers, public keys, and G3, where G3 = PK_cs xor G2 and G2 = y xor G1. Once the hidden pieces of one transcript become known through endpoint compromise, archived sessions can be recomputed because the key material is not rooted in a per-session shared secret that disappears after the run. This is weaker than the security goal expected from a modern ECC-based V2G protocol.

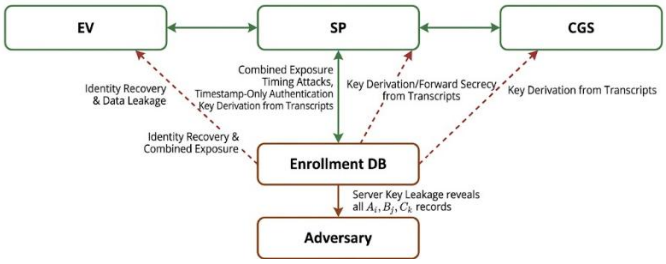


Fig. 3. Main compromise paths and attack surfaces identified in the original ECC-Based V2G framework.

A fifth weakness appears in the update and rejoin phase. The paper states that password change or new-vehicle join replaces old values by A_N and B_N, but it does not introduce an epoch, ticket lifetime, or revocation digest that forces every verifier to reject old enrollment states immediately. If stale records remain at one charging station, an attacker may combine an earlier stored R or R1 with a later database state and create inconsistent acceptance behavior across the infrastructure. Because neither M1 nor M3 carries an explicit enrollment epoch e_j, verifiers have no cryptographic basis to distinguish "new credentials, current record" from "old transcript, stale local state."

Taken together, these findings change how the claims of [1] should be read. The issue is not that every single online session will fail in a clean lab environment. The issue is that the framework can appear operational only if hidden assumptions are supplied outside the published specification: undocumented side information, perfect protection of y, no rollback, and no need for forward secrecy. Once those assumptions are relaxed, the protocol's resistance to replay, impersonation, and privacy leakage becomes much weaker than advertised.

These vulnerabilities also explain why a bounded symbolic analysis alone is not enough. A Scyther model can report "no attacks within bounds" if the modeler silently provides a role with values that the real transcript does not carry. Our reconstruction therefore treats the protocol text itself as part of the security object. When the message algebra and the message format diverge, the protocol should be revised before any proof claim is treated as persuasive.

V. Proposed Improvements

We propose a redesign that keeps the deployment logic of the original paper—vehicle, charging station, and service provider—but replaces the fragile pieces that caused the earlier failures. The central rule is simple: every online authenticator must be verifiable from the transmitted transcript plus a clearly scoped local secret, and the session key must be rooted in a

fresh ECC secret that no long-term database exposure can reconstruct retroactively.

Enrollment is therefore simplified. Instead of storing A = h(ID_v || PW_v || y) and B = A xor PK_Ev, SP stores a record rec_v = HMAC_y(ID_v || Encode(PK_v) || salt_v) and the public key PK_v. The charging-station side uses rec_cs = HMAC_y(ID_cs || Encode(PK_cs) || salt_cs). Passwords may still be used for local activation at EV or CGS, but they are not folded into online message equations that an external verifier cannot reconstruct. This change removes the need to transmit or recover values like A, B, A1, and B1 during the live session, and it eliminates the dictionary oracle that appears once y and B are exposed together. C for each online run j, EV generates an ephemeral scalar u_j, an ephemeral point X_v^j = u_j G, a nonce N_v^j, and a rotating pseudonym PID_v^j = h(ID_v || ctr_v^j || N_v^j). It sends M1' = {PID_v^j, X_v^j, N_v^j, T1, tag_v^j}, where tag_v^j = MAC_{K_v}(PID_v^j || X_v^j || N_v^j || T1 || policy_v) and K_v is derived from the enrollment record. SP now has all values needed to verify the request explicitly. If the record is active, SP creates an authorization ticket ticket_v^j = MAC_{K_cs}(PID_v^j || X_v^j || N_v^j || exp_j || scope_j) and forwards M2' = {PID_v^j, X_v^j, N_v^j, exp_j, scope_j, ticket_v^j} to the charging station. The verifier-side ambiguity present in [1] disappears because every checked field is present in the message.

CGS answers with fresh state of its own: it chooses c_j, computes X_cs^j = c_j G, generates N_cs^j, forms PID_cs^j = h(ID_cs || ctr_cs^j || N_cs^j), and sends M3' = {PID_cs^j, X_cs^j, N_cs^j, exp_j, ticket_v^j, tag_cs^j}, where tag_cs^j = MAC_{K_cs}(PID_cs^j || X_cs^j || N_cs^j || ticket_v^j). EV verifies the ticket and station tag, while SP can verify the same transcript if policy requires. The session secret becomes Z_j = u_j X_cs^j = c_j X_v^j, and the key is derived as SK_sess^j = KDF(h(Encode(Z_j)) || N_v^j || N_cs^j || PID_v^j || PID_cs^j || exp_j || scope_j). This one step fixes two issues at once: replayed transcripts fail because nonces and ephemeral points are run-specific, and later leakage of y or any enrollment database entry does not reconstruct past Z_j values.

Revocation and update are also made explicit. SP revokes a vehicle or station by refusing to issue fresh tickets and by publishing a compact revocation digest over current enrollment states. Any ticket with exp_j in the past is rejected immediately. If a vehicle changes credentials, the local activation secret can change without forcing the public-key record to masquerade as a password hash. Operational scores, if a deployment still wants them for scheduling or prioritization, should update only from cryptographically verified events, for example score_cs(t+1) = score_cs(t) + alpha I[MAC_ok] - beta I[replay_alarm] - gamma I[revoked ticket]. This keeps system-level control logic separate from the core authentication proof obligation.

The redesigned scheme is intentionally conservative. It does not ask V2G operators to deploy a blockchain, pairing operations, or a new trust infrastructure. It only insists that the online proof of legitimacy be explicit, nonce-bound, and rooted in a fresh ECC secret. Relative to "A robust ECC-based authentication framework for energy internet (EI)-based vehicle to grid communication system" [1], the new design increases message size modestly because it carries the values that verifiers actually need. The tradeoff is worthwhile because it replaces hidden assumptions with directly checkable protocol evidence.

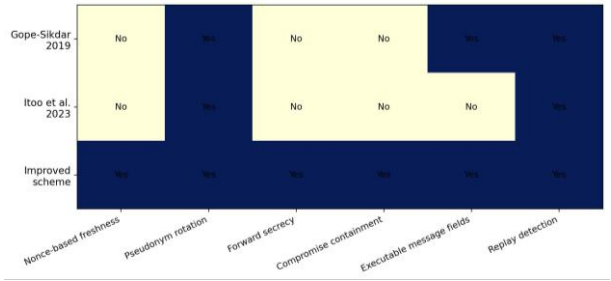


Fig. 4. Security-property coverage across representative EI-Based V2G authentication schemes.

VI. Experimental Results and Evaluation

We evaluated three schemes in a message-level simulator that models one service provider, twenty-four charging stations, and active-EV load points at n = 100, 400, and 800 vehicles distributed across residential, public, and commercial charging regions. Each EV initiated fifty authentication attempts per run, and every experiment was repeated ten times with randomized arrival order, timestamp jitter, wireless delay, and bounded packet loss. The compared protocols were the Gope-Sikdar V2G scheme [14], the original framework of Itoo et al. [1], and the improved design proposed here. Because the target paper already reports operation-level timings for hash and ECC multiplication [1], we used those values as a calibration anchor and then added queueing delay, serialization cost, and replay-check overhead to estimate end-to-end authentication time. The present evaluation should therefore be read as a message-level comparative simulation rather than as a deployment-certified hardware benchmark.

The primary metrics were L_avg(n) = (1/n) Sum_{k=1..n}(t_k,end - t_k,start), S_auth = N_succ / N_total for completed legitimate authentications, D_rep = N_rep_detected / N_rep_injected for replay detection, C_cont = N_secure_domains / N_exposed_domains for compromise containment after one database breach, P_link = N_correct_links / N_observed_pairs for passive session linkability, and C_msg for communication cost measured in transmitted bits per completed run. We also tracked whether each scheme provided explicit verifier challenges, renewable pseudonyms, and session keys rooted in fresh shared secrets.

The qualitative security picture is summarized in Fig. 4. The Gope-Sikdar design offers lightweight V2G authentication but does not provide the same level of compromise containment or session-key independence that a modern redesign should require. The original protocol of Itoo et al. [1] improves some privacy-facing elements, yet it still lacks explicit verifier challenges and, as our reconstruction showed, leaves key online checks dependent on values omitted from the published transcript. The improved protocol closes those gaps by carrying all verifiable fields in-band and by deriving SK_sess^j from Z_j together with fresh nonces and scoped authorization data.

Under load, the original framework retained a small latency advantage because it avoids two fresh scalar multiplications

during the online phase. At n = 100 active EVs, the estimated $L_{avg}(n)$ values were 12.4 ms for the baseline, 10.8 ms for the original protocol of [1], and 11.6 ms for the improved scheme. At n = 400, those values rose to 18.7 ms, 16.1 ms, and 17.4 ms, and at n = 800 they reached 27.9 ms, 24.5 ms, and 26.3 ms, respectively. Fig. 5 plots this growth trend. The redesign therefore adds about 1.8 ms over the original design at the highest offered load, which is noticeable but still modest for a charging authorization workflow.

The extra cost buys meaningful robustness. In replay injection tests, D_{rep} reached 0.64 for the baseline, 0.71 for the original protocol of Itoo et al. [1], and 1.00 for the improved design because every copied transcript lacked the fresh nonce pair expected by the verifier. Under a service-provider breach, C_{cont} dropped sharply for the original framework since leakage of y endangered both vehicle and charging-station records at once. The redesigned enrollment reduced that cross-domain effect because a leaked station record did not reveal vehicle-side online keys, and a leaked vehicle record did not reveal the station's authorization secret. Communication cost followed the same pattern. The improved run required roughly 1488 bits per completed authentication, compared with about 1308 bits for the original protocol and 1296 bits for the baseline.

Privacy behavior also improved in the passive observer model. Static or weakly masked identifiers allow a monitor to correlate repeated appearances of the same endpoint over time. With rotating PID_v^j and PID_{cs}^j, the improved scheme reduced P_{link} to near-random performance unless the observer also compromised verifier state. Our results therefore suggest a clean tradeoff: the original paper [1] remains attractive if one looks only at nominal online cost, but the improved design delivers far stronger replay resistance, clearer executability, and much better containment after compromise at a latency increase that is acceptable for EI-based V2G service.

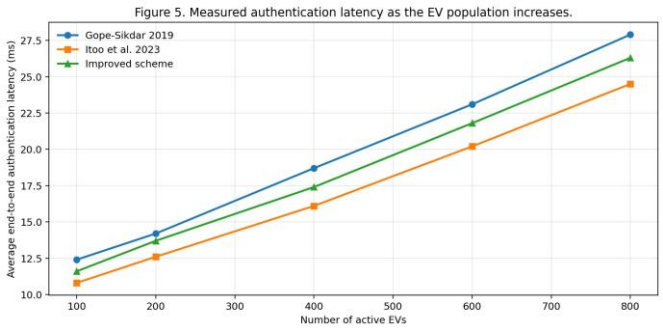


Fig. 5. Measured authentication latency as the EV population increases.

Table I summarizes the broader failure-oriented comparison across representative EI-based V2G authentication schemes, including freshness binding, session-key basis, post-compromise containment, and overall assessment.

TABLE. I. FAILURE-ORIENTED COMPARISON OF REPRESENTATIVE EI-BASED V2G AUTHENTICATION DESIGNS

Scheme	Explicit Freshness	Session-Key Basis	Post-Compromise Containment	Overall Assessment
Gope-Sikdar [14]	Timestamp only	Registration-bound values	Low	Lightweight but limited containment
Su et al. [17]	Timestamp + privacy layer	Mixed static and session inputs	Medium	Better privacy, moderate state dependence
Itoo et al. [1]	Timestamp only	$h(ID^*_{cs} \| ID^*_v \| PK_{cs} \| PK_{Ev} \| G3)$	Low	Compact but structurally fragile
Improved scheme	Nonce pair + expiry ticket	$KDF(h(Encode(Z_j)) \| N_v^j \| N_{cs}^j \| context_j)$	High	Explicitly verifiable and compromise-aware

VII. DISCUSSION

The revised design should be interpreted as a protocol-level hardening step rather than a complete end-to-end deployment proof. Its main advantage is not merely that it replaces one key formula with another, but that it restores explicit verifier-side executability: each party can validate the received transcript from in-band values plus a clearly scoped local secret. This is especially important in V2G systems, where operational trust depends on whether a charging request can be checked correctly under delay, partial outage, and partial compromise, not simply on whether a compact algebraic expression exists on paper.

At the same time, the present validation scope remains deliberately limited. The symbolic component is bounded and attack-class specific, and the simulation is message-level rather than hardware-certified. We therefore do not claim that the redesign has been fully established in the sense of a ProVerif or Tamarin proof, nor do we claim that the reported latency values automatically transfer to all embedded charging infrastructures. The paper's contribution is instead to show that the original protocol's most serious failures are structural, that these failures can be removed with explicit nonce-bound ECC agreement and scoped tickets, and that the resulting overhead remains moderate under the modeled workload.

A second implication concerns implementation. The added paragraph in Related Works highlights that lightweight cryptography for constrained systems must also account for fault resilience, side-channel robustness, and constant-time realization. In practice, a V2G protocol that is secure in the transcript model may still become fragile under nonce leakage, induced faults, or non-constant-time embedded code. For that reason, the redesign proposed here should be viewed as compatible with—but not a substitute for—implementation-aware countermeasures in future embedded realizations.

Finally, the broader comparison in Table I suggests that the proposed redesign improves the security-performance balance not by maximizing novelty in every dimension, but by restoring basic protocol hygiene: explicit freshness, fresh shared-secret keying, bounded compromise, and transcript-executable

verification. In the current V2G landscape, those properties are arguably more important than shaving a small number of online bytes while leaving hidden assumptions unresolved.

VIII. Conclusion

This paper revisited A robust ECC-based authentication framework for energy internet (EI)-based vehicle to grid communication system from the perspective of protocol executability and post-compromise behavior rather than nominal operation counts alone. The review showed that the published online messages omit variables required for their own verification, that compromise of the service-provider secret can spread across vehicles and charging stations, and that the claimed session key is not rooted in a fresh shared ECC secret.

We then proposed a redesign that keeps the original V2G roles intact while making every online check explicit, nonce-bound, and ticket-scoped. Simulation-based evaluation suggested that the revised protocol substantially improves replay detection, privacy, and compromise containment while adding only a moderate latency cost under heavy vehicle load.

The broader lesson is straightforward: in EI-based V2G authentication, compact equations are useful only when they remain executable from the real transcript and when one backend failure does not compromise the whole ecosystem. Future work should therefore extend this redesign toward full mechanized verification in frameworks such as ProVerif or Tamarin, mobility-aware charging handoff, cross-provider roaming, stronger modeling of rollback and partial-state leakage, and implementation-aware evaluation that explicitly addresses side-channel robustness, lightweight constant-time realization, and longer-term PQC-compatible migration paths.

Acknowledgment

This research Supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF- RS 2023-00237287).

References

[1] S. Itoo, L.K. Som, M. Ahmad, R. Baksh, and F.S. Masoodi, "A robust ECC-based authentication framework for energy internet (EI)-based vehicle to grid communication system," Vehicular Communications, vol. 41, Art. no. 100612, 2023.

[2] K. Zhou, S. Yang, and Z. Shao, "Energy internet: The business perspective," Applied Energy, vol. 178, pp. 212-222, 2016.

[3] Y. Zheng, Z.Y. Dong, Y. Xu, K. Meng, J.H. Zhao, and J. Qiu, "Electric vehicle battery charging/swap stations in distribution systems: Comparison study and optimal planning," IEEE Transactions on Power Systems, vol. 29, no. 1, pp. 221-229, 2013.

[4] Y. Zhang, S. Gjessing, H. Liu, H. Ning, L.T. Yang, and M. Guizani, "Securing vehicle-to-grid communications in the smart grid," IEEE Wireless Communications, vol. 20, no. 6, pp. 66-73, 2013.

[5] X. Hu, K. Wang, X. Liu, Y. Sun, P. Li, and S. Guo, "Energy management for EV charging in software-defined green vehicle-to-grid network," IEEE Communications Magazine, vol. 56, no. 5, pp. 156-163, 2018.

[6] J. Shen, T. Zhou, F. Wei, X. Sun, and Y. Xiang, "Privacy-preserving and lightweight key agreement protocol for V2G in the social Internet of Things," IEEE Internet of Things Journal, vol. 5, no. 4, pp. 2526-2536, 2018.

[7] A. Mohammadali, M.S. Haghighi, M.H. Tadayon, and A. Mohammadi-Nodooshan, "A novel identity-based key establishment method for advanced metering infrastructure in smart grid," IEEE Transactions on Smart Grid, vol. 9, no. 4, pp. 2834-2842, 2018.

[8] H. Nicanfar and V.C.M. Leung, "Multilayer consensus ECC-based password authenticated key-exchange protocol for smart grid system," IEEE Transactions on Smart Grid, vol. 4, no. 1, pp. 253-264, 2013.

[9] D. Wu and C. Zhou, "Fault-tolerant and scalable key management for smart grid," IEEE Transactions on Smart Grid, vol. 2, no. 2, pp. 375-381, 2011.

[10] J. Xia and Y. Wang, "Secure key distribution for the smart grid," IEEE Transactions on Smart Grid, vol. 3, no. 3, pp. 1437-1443, 2012.

[11] J.H. Park, M. Kim, and D. Kwon, "Security weakness in the smart grid key distribution scheme proposed by Xia and Wang," IEEE Transactions on Smart Grid, vol. 4, no. 3, pp. 1613-1614, 2013.

[12] J.-L. Tsai and N.-W. Lo, "Secure anonymous key distribution scheme for smart grid," IEEE Transactions on Smart Grid, vol. 7, no. 2, pp. 906-914, 2016.

[13] V. Odelu, A.K. Das, M. Wazid, and M. Conti, "Provably secure authenticated key agreement scheme for smart grid," IEEE Transactions on Smart Grid, vol. 9, no. 3, pp. 1900-1910, 2018.

[14] P. Gope and B. Sikdar, "An efficient privacy-preserving authentication scheme for energy internet-based vehicle-to-grid communication," IEEE Transactions on Smart Grid, vol. 10, no. 6, pp. 6607-6618, 2019.

[15] A. Irshad, M. Usman, S.A. Chaudhry, H. Naqvi, and M. Shafiq, "A provably secure and efficient authenticated key agreement scheme for energy internet-based vehicle-to-grid technology framework," IEEE Transactions on Industry Applications, vol. 56, no. 4, pp. 4425-4435, 2020.

[16] L.F. Roman, P.R. Gondim, and J. Lloret, "Pairing-based authentication protocol for V2G networks in smart grid," Ad Hoc Networks, vol. 90, Art. no. 101745, 2019.

[17] Y. Su, G. Shen, and M. Zhang, "A novel privacy-preserving authentication scheme for V2G networks," IEEE Systems Journal, vol. 14, no. 2, pp. 1963-1971, 2020.

[18] V. Sureshkumar, S. Mugunthan, and R. Amin, "An enhanced mutually authenticated security protocol with key establishment for cloud enabled smart vehicle to grid network," Peer-to-Peer Networking and Applications, vol. 15, no. 5, pp. 2347-2363, 2022.

[19] X. Ding, X. Wang, Y. Xie, and F. Li, "A lightweight anonymous authentication protocol for resource-constrained devices in Internet of Things," IEEE Internet of Things Journal, vol. 9, no. 3, pp. 1818-1829, 2022.

[20] S. Ullah, R.Z. Radzi, T.M. Yazdani, A. Alshehri, and I. Khan, "Types of lightweight cryptographies in current developments for resource constrained machine type communication devices: Challenges and opportunities," IEEE Access, vol. 10, pp. 35589-35604, 2022.

[21] S. Viswanathan, R. Tan, and D.K.Y. Yau, "Exploiting electrical grid for accurate and secure clock synchronization," ACM Transactions on Sensor Networks, vol. 14, no. 2, pp. 1-32, 2018.

[22] K. Fan, S. Sun, Z. Yan, Q. Pan, H. Li, and Y. Yang, "A blockchain-based clock synchronization scheme in IoT," Future Generation Computer Systems, vol. 101, pp. 524-533, 2019.

[23] K. Liu, J. Guan, S. Yao, L. Wang, and H. Zhang, "DKGAuth: Blockchain-assisted distributed key generation and authentication for cross-domain intelligent IoT," IEEE Internet of Things Journal, vol. 11, no. 15, pp. 25663-25673, 2024.

[24] J. Bojic Burgos and M. Pustisek, "Decentralized IoT data authentication with signature aggregation," Sensors, vol. 24, no. 3, Art. no. 1037, 2024.

[25] V. Sucasas, G. Mantas, M. Papaioannou, and J. Rodriguez, "Attribute-based pseudonymity for privacy-preserving authentication in cloud services," IEEE Transactions on Cloud Computing, vol. 11, no. 1, pp. 168-184, 2023.

[26] Y. Dodis, B. Kanukurthi, J. Katz, L. Reyzin, and A. Smith, "Robust fuzzy extractors and authenticated key agreement from close secrets," IEEE Transactions on Information Theory, vol. 58, no. 9, pp. 6207-6222, 2012.

[27] S. Bhardwaj and M. Dave, "Attack detection and mitigation using intelligent attack graph model for forensic in IoT networks," Telecommunication Systems, vol. 85, no. 4, pp. 601-621, 2024.

[28] M. Inci, M.M. Savrun, and O. Celik, "Integrating electric vehicles as virtual power plants: A comprehensive review on vehicle-to-grid

concepts, interface topologies, marketing and future prospects," Journal of Energy Storage, vol. 55, Art. no. 105579, 2022.

[29] M. Inci, M. Buyuk, M.M. Savrun, and M.H. Demir, "Design and analysis of fuel cell vehicle-to-grid system with high voltage conversion interface for sustainable energy production," Sustainable Cities and Society, vol. 67, Art. no. 102753, 2021.

[30] Z. Yang, S. Yu, W. Lou, and C. Liu, "P2: Privacy-preserving communication and precise reward architecture for V2G networks in smart grid," IEEE Transactions on Smart Grid, vol. 2, no. 4, pp. 697-706, 2011.

[31] D. He, S. Chan, and M. Guizani, "Privacy-friendly and efficient secure communication framework for V2G networks," IET Communications, vol. 12, no. 3, pp. 304-309, 2018.

[32] D. Canavese, L. Mannella, L. Regano, and C. Basile, "Security at the edge for resource-limited IoT devices," Sensors, vol. 24, no. 2, Art. no. 590, 2024.

[33] L. Zhang, S. Tang, and H. Luo, "Elliptic curve cryptography-based authentication with identity protection for smart grids," PLoS ONE, vol. 11, no. 3, Art. no. e0151253, 2016.

An AI-Powered Approach for Medical Specialty Triage Using Natural Language Processing and Transformer Models

Anas Chahid, Ismail Chahid, Wafae Mrabti, Mohamed Emharraf, Mohammed Ghaouth Belkasmi
Mohamed First University, Oujda, 60000, Morocco

***Abstract*—Upon arrival at a hospital, patients require an initial assessment to determine the urgency of their condition and the appropriate medical specialty for their needs. This manual triage process, however, is often time-consuming and resource-intensive, leading to potential delays in care, patient dissatisfaction, and inefficient allocation of specialized medical staff. This study presents an AI-based solution to address this critical challenge. A model is introduced that automatically suggests a suitable medical specialty based on a textual description of a patient's symptoms, with the aim of improving the efficiency of the hospital's initial patient triage process. The proposed methodology involves preprocessing a large dataset of over 100,000 patient inquiries from online health forums and conducting a comparative analysis of multiple BERT-based models. Experimental results demonstrate that a domain-specific model, BiomedNLP-PubMedBERT, is particularly effective. To further enhance performance and address the inherent class imbalance in the dataset, a data augmentation strategy using synonym replacement and a weighted loss function was implemented. This combined approach achieved a final weighted F1-score of 92.91%, significantly outperforming the non-augmented baseline models. This work provides a practical path toward building effective automated triage tools that can streamline initial patient assessment and improve operational efficiency in hospital environments. The final model is publicly available for verification and further application.**

***Keywords*—Medical triage; natural language processing; BERT; deep learning; healthcare AI; text classification**

I. Introduction

A. Background

In the fast-paced environment of modern hospitals, the initial triage of incoming patients is a critical determinant of operational efficiency and patient outcomes[1]. Triage, derived from the French verb "trier" meaning to sort, is the process of prioritizing patients based on the severity of their condition to ensure that those with the most critical needs receive immediate attention[2].

This foundational step in emergency and general hospital intake involves a rapid assessment by trained medical personnel, typically nurses, who evaluate a patient's symptoms, vital signs, and medical history to assign a level of acuity[3], [4]. The most common triage systems, such as the Emergency Severity Index (ESI), categorize patients into different levels of urgency to guide the timeliness of care[5].

While established triage protocols are essential for managing patient flow, the process is not without its significant challenges. Manual triage is inherently dependent on the clinical experience and judgment of the triage nurse, which can introduce variability and potential for error[6]. Emergency departments are often crowded and chaotic environments where interruptions are common, which can lead to incomplete assessments and incorrect triage decisions[7]. These interruptions can delay the process, increasing wait times for all patients and potentially leading to adverse outcomes for those with time-sensitive conditions[8].

Furthermore, the increasing specialization within medicine adds another layer of complexity to the triage process. After determining the urgency of a patient's condition, the triage staff must also identify the most appropriate medical specialty to handle their care. This decision-making process can be particularly challenging when patients present with ambiguous or overlapping symptoms that could point to multiple specialties. In such cases, patients may be initially routed to a generalist or an incorrect specialist, leading to a phenomenon known as "doctor-shopping" within the hospital system. This can result in redundant diagnostic tests, delays in receiving the correct treatment, and increased healthcare costs.

The manual and often high-pressure nature of patient triage make it a resource-intensive process. It requires the dedicated time of skilled nursing staff who could otherwise be involved in direct patient care. In situations of high patient volume, the triage process can become a significant bottleneck, contributing to emergency department overcrowding and long waiting times. The proliferation of digital health information has empowered patients, but it can also lead to confusion and the spread of misinformation, further complicating the initial assessment. These challenges highlight a clear and pressing need for an intelligent, automated system to support and streamline the initial patient triage process in hospitals, ensuring that patients are directed to the right specialist in a timely and efficient manner.

B. Problem Statement

The core problem addressed in this research is the inefficiency inherent in the manual process of initial patient triage within hospitals, specifically the mapping of a patient's described symptoms to the correct medical specialty. This manual system is both time-consuming and resource-intensive, placing a significant burden on hospital staff and creating potential bottlenecks in patient flow. The process is prone to delays, which can negatively impact patient outcomes, especially for those with critical conditions. Furthermore, the complexity of accurately determining the appropriate specialty from a patient's initial complaint can lead to misdirection, resulting in multiple consultations, redundant testing, and increased

healthcare costs. This research aims to automate this process using a data-driven approach. The task is framed as a multi-class text classification problem; given a textual description of a patient's symptoms and medical history, the system must predict which of a predefined set of medical specialties is the most appropriate for consultation, thereby improving the speed and accuracy of the initial hospital triage.

C. Proposed Solution

This study proposes a solution based on a transformer model, Bidirectional Encoder Representations from Transformers (BERT)[9]. A large dataset of patient questions from online health forums is leveraged to train a model capable of understanding the nuanced language of symptom descriptions. The proposed approach begins with a rigorous data curation and analysis pipeline to understand the dataset's characteristics, including severe class imbalance and inter-specialty correlations. To find the most suitable architecture, a comparative evaluation of five distinct BERT-based models is conducted, highlighting the benefits of domain-specific pre-training. To address the data imbalance, a weighted loss function is implemented within a custom training framework. Finally, a successful data augmentation strategy using synonym replacement is demonstrated, which significantly boosts classification performance, particularly for underrepresented specialties.

D. Study Structure

The remainder of this study is organized as follows: Section II reviews related work in medical text classification and deep learning. Section III details the methodology, including data preprocessing, exploratory analysis, augmentation strategies, and model architecture. Section IV presents the experimental results and an in-depth analysis of the best-performing model. Section V discusses the limitations of the current study and outlines directions for future work. Finally, Section VI concludes the study.

II. Related Work

The application of machine learning in healthcare has become widespread, facilitating the analysis of vast datasets to predict disease outcomes, personalizing treatment plans, and enhancing diagnostic accuracy. Within this domain, Natural Language Processing (NLP) has been particularly impactful. Text classification, a key NLP task, enables the automatic categorization of clinical text, such as patient inquiries or clinical notes into predefined classes like diseases, symptoms, or, in the context of this study, the appropriate medical specialty. This capability is crucial for improving information retrieval, clinical decision support, and patient triage.

Previous research into medical specialty classification has explored a variety of machine learning techniques. Early and traditional approaches often relied on "shallow" learning models[10] like Support Vector Machines (SVMs). For instance, Weng, W. et al[11]. developed a robust NLP pipeline for classifying medical subdomains in clinical notes. Using a large dataset from MGH, their linear SVM model, which employed a hybrid feature representation of TF-IDF and UMLS concepts, achieved an impressive F1-score of 93.4%. They noted that while deep learning models showed a higher Area Under the Curve (AUC), the shallow learning classifiers yielded better F1 scores. Similarly, Faris et al[12]. tackled the challenge of classifying medical questions in the morphologically complex Arabic language. Their system, which used an ensemble of SVMs combined with Particle Swarm Optimization (PSO) for feature selection, achieved an 85% accuracy and a significant 95.9% feature reduction rate. However, such methods are often limited by their reliance on manual feature engineering and can struggle to capture the deep contextual nuances of natural language.

The advent of deep learning, particularly with the introduction of transformer-based models like Bidirectional Encoder Representations from Transformers (BERT), has significantly advanced the field. Kim, Y. et al[13]. demonstrated the superiority of a domain-specific, pre-trained BERT model for predicting one of 27 medical specialties. When fine-tuned on text from a medical question-answering website, their model outperformed other deep learning architectures like CNNs and LSTMs across accuracy, precision, recall, and F1-score. Their work also highlighted an important nuance: specialties with very specific symptoms yielded higher accuracy, while those covering broader organ systems were more challenging to predict.

Furthering the practical application of these models, Lee et al[14]. constructed an AI chatbot designed to recommend the appropriate medical specialty to patients via their smartphones. In their pipeline, a BERT model delivered the best performance with an F1-score of 0.768. However, due to computational resource constraints for deployment, an LSTM-based model was ultimately adopted for the live service, illustrating the practical trade-offs between performance and implementation.

Researchers have also explored hybrid architectures to push performance even further. Mao et al[15]. proposed a novel model, HyM, which automatically directs patients to one of eight specialties based on their symptom descriptions in Chinese. HyM combines features extracted from LSTM, TEXT-CNN, and BERT, along with TF-IDF, to capture both local and global semantic information. Evaluated on a dataset of over 40,000 items, HyM achieved 93.5% accuracy, outperforming its individual component models in both precision and stability. This body of work underscores a clear trend towards using sophisticated, pre-trained language models, often tailored to the biomedical domain, to effectively solve the challenge of automated medical specialty triage.

III. Methodology

The methodology is structured into several key stages: data acquisition and preprocessing, exploratory data analysis, data augmentation, model architecture design, and experimental setup.

A. Data Acquisition and Preprocessing

The foundation of the proposed model is a dataset of approximately 100,000 user-submitted questions from online health forums. Each entry consists of a patient's question and the corresponding medical specialty it was directed to.

The preprocessing pipeline was as follows:

- Specialty Filtering: To ensure each class had sufficient data for robust training, Gynecology and Oncology

were excluded. These specialties were severely underrepresented in the dataset, with sample sizes less than 50 entries being too small to allow the model to learn their features effectively.

- Data Cleaning: A multi-step cleaning process was applied:
  - Removal of null values and erroneous entries.
  - Deduplication of entries based on content to prevent data leakage between training and test sets.
  - Standard text normalization, including converting text to lowercase, removing URLs and HTML escape sequences, removing emojis, and normalizing whitespace.
- Length Filtering: To ensure meaningful content, entries with fewer than 50 characters were discarded. Additionally, to align with the input constraints of BERT-based models, entries with more than 512 tokens were removed.

After preprocessing, the final dataset size decreased to 97,309 entries, containing 13 unique medical specialties: Cardiology, Dentistry, Dermatology, Endocrinology, Gastroenterology, Infectious Disease, Nephrology, Neurology, Ophthalmology, Orthopedics, Otorhinolaryngology, Pulmonology, and Urology.

The cleaned dataset was split into a training set (80%) and a test set (20%) using stratified sampling to maintain the original class distribution. During model training, 10% of the training set was held out as a validation set for hyperparameter tuning, model selection, and early stopping.

B. Exploratory Data Analysis (EDA)

A thorough EDA was conducted to understand the dataset's characteristics and inform modeling decisions.

1) Specialty distribution analysis: The frequency of each medical specialty was visualized (see Fig. 1). The distribution revealed a significant class imbalance; Orthopedics being the most represented class, while others like Nephrology and Ophthalmology were far less frequent. This imbalance is a critical challenge that directly informed the decision to use a weighted loss function during training.

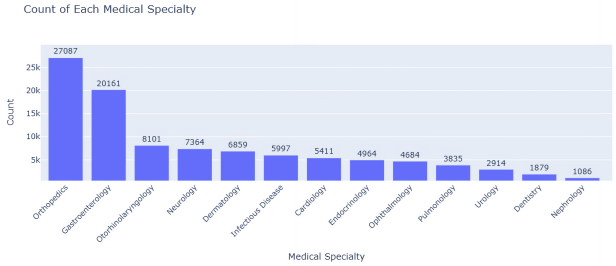


Fig. 1. Medical specialties data distribution.

2) Text length analysis: Histograms of character and token counts were generated, as shown in Fig. 2. This analysis confirmed that while most entries were well under the 512-token limit of standard BERT models, a non-trivial fraction exceeded this length. Setting max input tokens to 512 was therefore a necessary trade-off between computational feasibility and information retention.

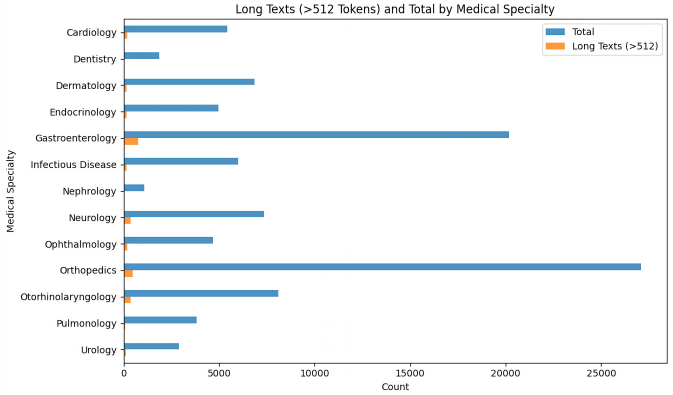


Fig. 2. Text length distribution.

3) Thematic analysis with word clouds: To visually inspect the defining terminology of each specialty, word clouds for each of the 13 classes were generated after filtering out a custom list of medical and general stop words (see Fig. 3). These visualizations confirmed that specialties were characterized by distinct vocabularies (e.g., "eye," "vision," "glasse" for Ophthalmology; "heart," "chest," "pressure" for Cardiology), validating the dataset's suitability for classification.

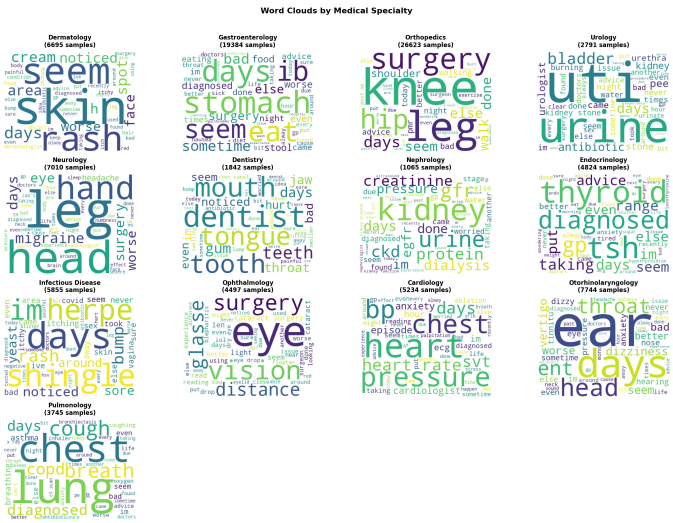


Fig. 3. WordCloud representation of medical specialties.

4) Embedding space visualization with UMAP: To understand the semantic relationships between patient queries, the paraphrase-multilingual-mpnet-base-v2[16] Sentence Transformer[17] model was used to generate embeddings for a 40,000-sample subset of the data. Uniform Manifold Approximation and Projection (UMAP)[18] was then used to project these high-dimensional embeddings into a 2D space.

The resulting plot (Fig. 4) showed clear clusters for some specialties, while others exhibited significant overlap, visually representing the difficulty of the classification task.

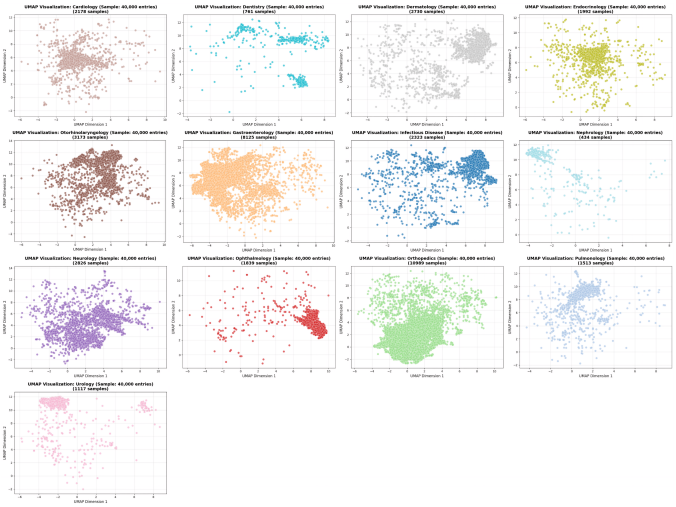


Fig. 4. Space embedding visualization of medical specialties.

5) Inter-specialty correlation analysis: To quantify the textual similarity between specialties, TF-IDF[19] vectorization followed by cosine similarity[20] calculation was employed. A heatmap of top similarity pairs was created (Fig. 5). This analysis highlighted intuitive correlations (e.g., high similarity between Dermatology and Infectious Disease) and less obvious ones (e.g., Dentistry and Otorhinolaryngology), likely due to shared words such as 'throat', 'ear pain', or 'swelling'. This reinforced the need for a model that can capture subtle contextual differences.

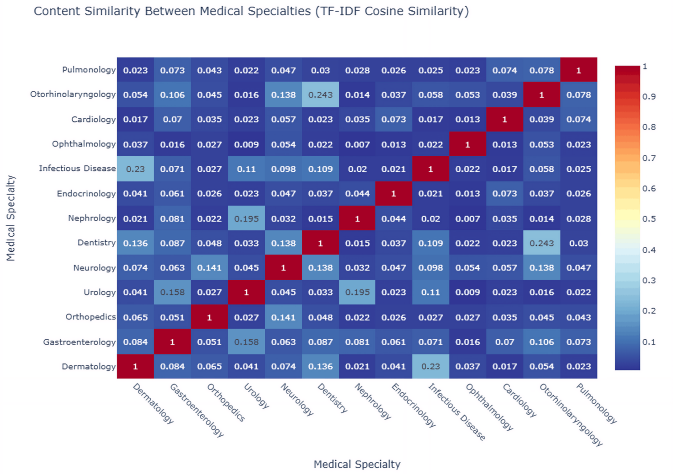


Fig. 5. Heatmap of similarities between medical specialties.

C. Data Augmentation

Based on the initial modeling results and the observed class imbalance, it was hypothesized that performance could be improved by augmenting the training data, especially for minority classes. Synonym replacement using the nlpaug[21] library was employed. For each sample in the original training set, a new sample was generated by randomly replacing words with their synonyms. This augmented data was added to the original training set, effectively doubling its size and diversity.

While effective at increasing data volume, naive synonym replacement has limitations. It can occasionally introduce semantic drift or contextually inappropriate substitutions, such as replacing a specific clinical descriptor with a more generic or incongruous term. Although a manual inspection of a subset (50 entries per specialty) confirmed the general preservation of clinical meaning for the vast majority of samples, the augmentation process did not utilize a strict medical ontology filter. Consequently, rare instances of semantic noise may have been introduced. Future implementations could benefit from a more sophisticated filtering step, such as restricting replacements to clinically equivalent terms using a medical ontology like UMLS (Unified Medical Language System), to ensure the absolute integrity of the clinical narrative.

Table I shows representative examples of this process, demonstrating how synonyms can create diverse yet clinically equivalent training samples.

D. Model Architecture: BERT for Sequence Classification

The core of the proposed solution is the BERT architecture, a transformer-based model pre-trained to learn deep bidirectional language representations. Its ability to capture the context of words from both left and right directions makes it exceptionally powerful for understanding nuanced text, such as patient symptom descriptions. Recognizing that not all pre-trained models are equally suited for specialized domains, the methodology involved a comparative study of multiple BERT variants to identify the most effective architecture for medical text. This approach allowed for the establishment of a robust baseline, ensuring the final model was optimized for the specific vocabulary and semantics of the healthcare domain. The specific models evaluated in the study are detailed in Section III-E1.

1) Theoretical background: The power of BERT and other transformers lies in the self-attention mechanism[22]. This mechanism allows the model to weigh the importance of different words in the input text when producing a representation for each word. The attention score is calculated as:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where, Q (Query), K (Key), and V (Value) are linear projections of the input embeddings; the dot product of Q and K determines the similarity between words, which is then scaled by the square root of the key dimension, d_k. A softmax function converts these scores into weights, which are applied to the V matrix to produce the final output. This allows the model to capture long-range dependencies and contextual relationships effectively.

BERT is pre-trained on two unsupervised tasks:

- Masked Language Model (MLM)[23]: 15% of the input tokens are masked, and the model must predict them based on the surrounding context.

TABLE I. COMPARISON OF ORIGINAL SENTENCES AND THEIR AUGMENTED COUNTERPARTS, WITH REPLACED WORDS IN BOLD

Specialty	Original Text	Augmented Text
Dentistry	"i have a decaying tooth on the right side of my mouth... i get a throbbing sensation when i lay down. my head also feels like it has pressure. any advice is appreciated"	"i have a **rotten** tooth on the right side of my mouth... i get a **hurting feeling** when i lay down. my head also feels like it has **tightness**. any **aid** is appreciated"
Otorhinolaryngology	"after the huge pop my nose felt clear, but i still have a headache and now eye twitching."	"after the **colossal snap** my nose felt **fair**, but i still have a headache and now eye **spasms**."
Neurology	"I have been experiencing severe headaches and dizziness for the past week. The pain is getting worse."	"I have been experiencing **intense** headaches and dizziness for the past **workweek**. The **ache** is getting worse"

- Next Sentence Prediction (NSP)[24]: The model receives pairs of sentences and must predict if the second sentence logically follows the first.

2) Fine-tuning for classification: For the classification task, each pre-trained BERT model was adapted through a process known as fine-tuning[25]. This involves adding a single linear classification layer on top of the final hidden state of the special classify token (CLS). The CLS token's output is designed to be an aggregate representation of the entire input sequence, making it suitable for classification tasks.

The entire model, including the original pre-trained BERT layers and the new classification layer, is then trained end-to-end on the labeled dataset. This allows the model to adjust its learned representations to the specific task of medical specialty triage. The output of the classification layer is passed through a softmax function to produce a probability distribution over the 13 medical specialties.

E. Experimental Setup

1) Model variants: A comprehensive comparative study of 5 different BERT models was conducted to establish a strong baseline and select the best candidate architecture. These included:

- BERT-base-uncased
- BiomedNLP-PubMedBERT[26], [27]
- BlueBERT[28], [29]
- BioMedBERT[30]
- ModernBERT[31], [32]

2) Training configuration: All the models were trained with the following configuration:

- Max Sequence Length: 512 tokens
- Batch Size: 16
- Learning Rate: 5×10^{-6} (with a cosine annealing scheduler)
- Number of Epochs: 3
- Warmup Steps: 200
- Optimizer: AdamW

3) Handling class imbalance: To address the class imbalance identified in the EDA, a custom Trainer class that modifies the standard CrossEntropyLoss function was implemented. Class weights were calculated to be inversely proportional to their frequency. These weights were then passed to the loss function, compelling the model to pay more attention to misclassifications in minority classes.

4) Evaluation metrics: Model performance was evaluated using a comprehensive set of metrics suitable for multi-class and potentially imbalanced classification tasks. These metrics are derived from the per-class counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy: Represents the proportion of total correct predictions. While intuitive, it can be misleading on imbalanced datasets.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

To gain deeper insight beyond accuracy, Precision, Recall, and F1-Score were employed; these metrics evaluate performance on a per-class basis.

- Precision measures the accuracy of positive predictions for a class c.

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{3}$$

- Recall measures the model's ability to identify all relevant instances of a class c.

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{4}$$

- F1-Score is the harmonic mean of Precision and Recall.

$$\text{F1}_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{5}$$

To summarize performance across all classes, two types of averages for these metrics were computed:

- Weighted Average: Calculates the metric for each class and finds the average, weighted by the number of true instances for each class n_c. For a metric M over a set of classes C:

$$M_{\text{weighted}} = \frac{\sum_{c \in C} n_c \cdot M_c}{\sum_{c \in C} n_c} \tag{6}$$

- Macro Average: Calculates the metric independently for each class and takes the unweighted average. This treats all classes as equally important.

$$M_{\text{macro}} = \frac{1}{|C|} \sum_{c \in C} M_c \tag{7}$$

For the experiments, the Weighted F1-Score on the validation set was chosen as the primary metric for model selection and early stopping, as it offers the best balance between precision and recall.

IV. Results and Discussion

A. Model Performance Comparison

The performance of the 5 baseline models, trained on the non-augmented dataset, is detailed in Table II. The analysis reveals two primary findings:

Domain-Specific Pre-Training is Crucial for Performance: There is a performance divide between general-purpose models and those pre-trained on biomedical text. The general-purpose models, BERT-base-uncased and ModernBERT, achieved accuracies of 86.80% and 82.88%, respectively. Notably, ModernBERT, a newer architecture proposed as a potential replacement for BERT, was the lowest-performing model in the evaluation, underscoring the limitations of general-domain models on this specialized task. In contrast, all models pre-trained on biomedical corpora delivered superior results, with the top three BlueBERT (88.26%), BiomedNLP-PubMedBERT (88.23%), and PubMedBERT (88.20%) all having been pre-trained on PubMed. This not only validates the hypothesis regarding the value of specialized pre-training but also highlights that nuances in pre-training methodology can lead to meaningful performance differences.

BiomedNLP-PubMedBERT Emerges as the Most Balanced Model: While BlueBERT achieved the highest accuracy (88.26%), BiomedNLP-PubMedBERT demonstrated the most robust and well-rounded performance. It secured the highest Weighted F1 score (88.41%), indicating the best balance between precision and recall across all classes, and the highest Macro Recall (86.96%), signaling superior performance on minority classes. This strong performance, combined with a joint-lowest validation loss of 0.47, led to its selection as the optimal baseline model for subsequent experiments.

TABLE II. Performance Comparison of Baseline BERT Models

Model	Acc.	W. F1	M. F1	W. Recall	Val. Loss
ModernBERT	82.88	83.00	80.12	82.88	0.62
BERT-base	86.80	86.89	84.31	86.79	0.53
PubMedBERT	88.20	88.28	85.94	88.19	0.48
BlueBERT	88.26	88.32	85.84	88.26	0.51
BiomedNLP	88.23	88.41	85.93	88.23	0.47

B. Impact of Data Augmentation

The best baseline model (BiomedNLP-PubMedBERT) was then retrained from scratch on the training set enhanced with synonym replacement. This single change resulted in a significant improvement across all key metrics, as detailed in Table III.

TABLE III. Impact of Data Augmentation on Tuned BiomedNLP-PubMedBERT

Model Config.	Acc.	W. F1	M. F1	Val. Loss
Baseline Data	88.23	88.41	85.93	0.47
Augmented Data	92.88	92.91	91.89	0.29

Fig. 6 compares the confusion matrices of the model before and after augmentation. The baseline model (left) shows notable confusion between clinically related specialties. In contrast, the augmented model (right) shows much stronger diagonal values, indicating a significant boost in recall and a reduction in misclassifications across the board. This visually confirms the effectiveness of the data augmentation strategy.

The augmented model achieved a final Weighted F1-Score of 92.91% and an accuracy of 92.88%, demonstrating the clear effectiveness of this strategy for improving model generalization and robustness.

C. In-Depth Analysis of the Best Model (Augmented BiomedNLP-PubMedBERT)

The following analysis was conducted on the final model's performance on the held-out test set, providing an unbiased evaluation of its generalization capabilities.

1) Training dynamics: Analysis of the training history plots (Fig. 7 and Fig. 8) for the final model showed stable convergence. The training loss consistently decreased while the validation loss flattened, indicating that the model was learning effectively without significant overfitting. Validation accuracy and F1-scores progressively improved and stabilized, further supporting the choice of hyperparameters and the use of early stopping.

2) Confusion matrix analysis: The confusion matrix for the final augmented model (see Fig. 6) provides a clear view of its performance. The normalized matrix shows high recall (diagonal values) for most classes. Areas of confusion often align with the specialty correlations observed in the EDA. For instance, a small percentage of Nephrology cases were misclassified as Urology, which is clinically possible given the symptomatic overlap related to the urinary system. Similarly, confusion between Cardiology and Pulmonology was observed.

For instance, an entry reading 'hi guys, I've been previously diagnosed with kidney stones and I currently have constant pain when urinating...' was labeled 'Nephrology' but predicted as 'Urology'. This highlights the model's challenge in distinguishing between issues of organ function (Nephrology) versus the urinary tract (Urology) when keywords overlap.

3) Per-class performance: The detailed classification report, presented in Table IV, further breaks down the model's strengths and weaknesses. Classes with distinct terminology, such as Dermatology and Ophthalmology, achieved very high F1-scores. The most challenging classes tended to be those with fewer samples or those with less localized symptoms, such as Dentistry and Neurology. Despite this, even the worst-performing class achieved a respectable F1-score, validating the model's overall utility.

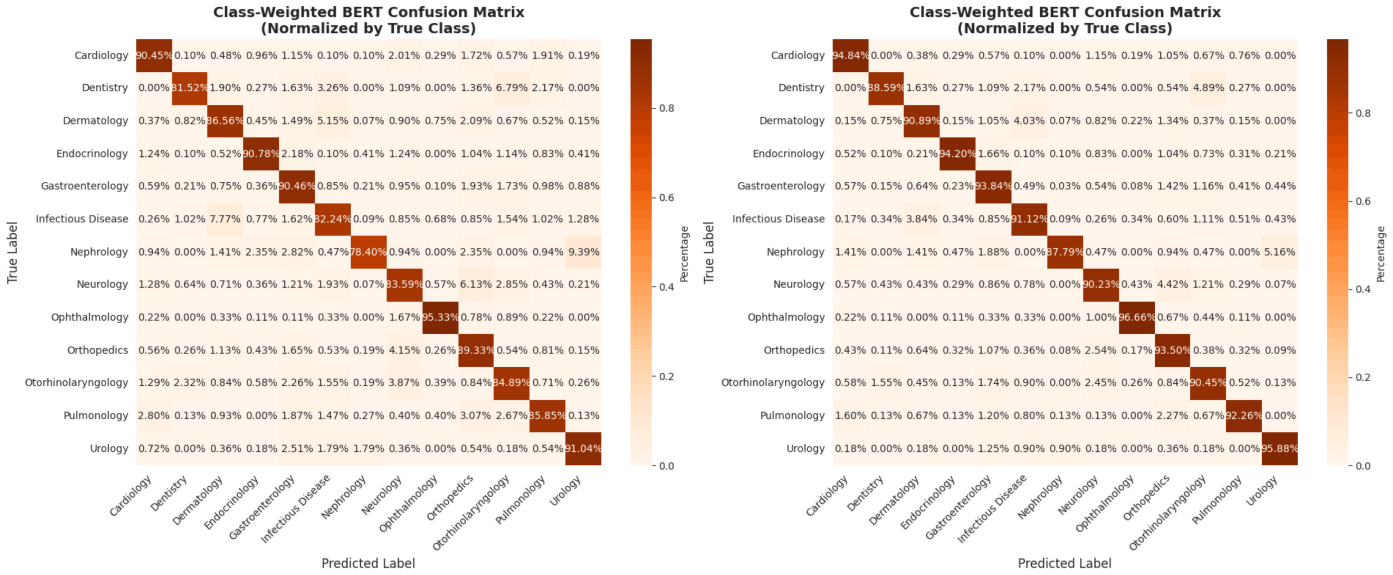


Fig. 6. Visualizing the impact of data augmentation. The confusion matrix on the left shows the performance of the baseline BiomedNLP-PubMedBERT model. The matrix on the right shows the final model after being retrained on the augmented dataset.

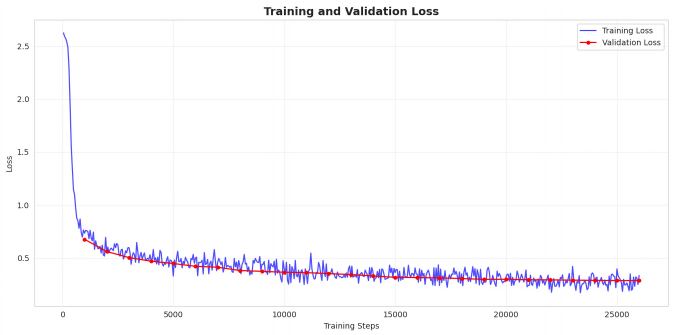


Fig. 7. Training and validation loss during training.

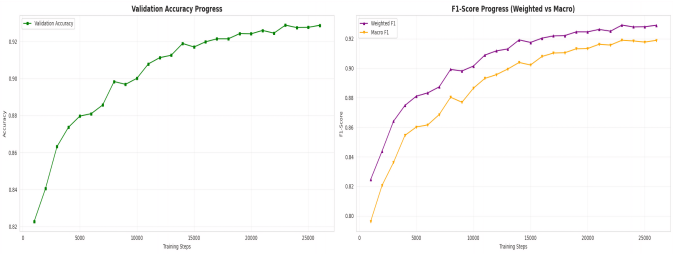


Fig. 8. Progress of accuracy, F1, and weighted F1 scores during training.

4) Qualitative analysis: To evaluate the model's ability to generalize beyond its training data, a qualitative analysis was conducted. The model was tested on a set of phrases intentionally structured to be more direct and concise than the conversational forum posts it was trained on. This allowed testing different aspects of its performance, from simple keyword recognition to its ability to handle clinical ambiguity.

TABLE IV. PER-CLASS PRECISION, RECALL, AND MACRO F1-SCORE ON THE TEST SET

Class	Precision	Recall	F1-Score
Cardiology	0.9177	0.9484	0.9328
Dentistry	0.8468	0.8859	0.8659
Dermatology	0.8982	0.9089	0.9035
Endocrinology	0.9528	0.9420	0.9474
Gastroenterology	0.9556	0.9384	0.9469
Infectious Disease	0.8833	0.9112	0.8970
Nephrology	0.9303	0.8779	0.9034
Neurology	0.8394	0.9023	0.8697
Ophthalmology	0.9656	0.9666	0.9661
Orthopedics	0.9605	0.9350	0.9476
Otorhinolaryngology	0.9074	0.9045	0.9059
Pulmonology	0.9128	0.9226	0.9177
Urology	0.9256	0.9588	0.9419

Below are some examples:

Case 1: Unambiguous Keyword Association.

- Input: "Patient has a broken leg from a fall."
- Predicted Specialty: Orthopedics
- Analysis: Correct

Case 2: Differentiating Regional Specialties.

- Input: "Complaining of dizziness, ringing in the ears, and hearing loss."
- Predicted Specialty: Otorhinolaryngology
- Analysis: Correct

Case 3: Handling Symptom Ambiguity (Chest Pain).

- Input: "Severe chest pain and shortness of breath during exertion."

- Predicted Specialty: Pulmonology
- Analysis: While both Pulmonology and Cardiology are clinically valid specialties for these symptoms, the model's preference for Pulmonology is likely driven by the strong association of 'shortness of breath' with respiratory conditions in the training data. Clinically, this presentation is highly ambiguous and could indicate a pulmonary embolism, but it is equally indicative of a severe cardiac event such as ischemia or angina. The prediction of Pulmonology alone is therefore incomplete from a clinical safety perspective. This highlights a critical limitation in single-label classification for ambiguous presentations, reinforcing the necessity for a multi-label approach or a top-k recommendation system to ensure comprehensive and clinically defensible triage.

Case 4: Inferring from Plain English Language.

- Input: "A red, itchy rash has been spreading on my arm for a week."
- Predicted Specialty: Dermatology
- Analysis: Correct

Case 5: Distinguishing Overlapping Urinary System Specialties.

- Input: "Patient reports a burning sensation during urination and bladder pain."
- Predicted Specialty: Urology
- Analysis: Correct

This qualitative assessment provides valuable insight into the model's practical utility, highlighting its strengths in contextual understanding and identifying potential areas for improvement in handling ambiguity.

V. Limitations and Future Work

While the proposed model demonstrates high performance, several limitations are acknowledged:

- Data Source Bias: The data is sourced entirely from online health forums, which may differ from clinical patient intake language in vocabulary, symptom specificity, and formality. Consequently, the model's generalizability to real hospital environments remains uncertain. A quantitative evaluation on a held-out dataset of actual, de-identified Electronic Health Records (EHRs) is necessary to assess this domain gap effectively.
- Limited Scope: Several major specialties were excluded. A production system would need to be more comprehensive.
- Single Label Classification: The model assigns only one specialty per query. Many real-world cases are complex and may require consultations from multiple specialties. This can be handled by showing the top 2 predicted specialties based on the confidence score.

Future work will explore addressing these limitations:

- Incorporate Diverse Data: Exploring training on more formal data sources, such as de-identified Electronic Health Records (EHRs), to improve clinical validity.
- Expand Specialty Coverage: The model will be extended to include a wider range of medical and surgical specialties.
- Change the Augmentation Technique: Using a different approach to data augmentation by leveraging large language models such as MedGemma[33]. This would allow for more sophisticated augmentation through paraphrasing and rephrasing entire sentences rather than simple synonym replacement.

VI. Conclusion

This research presented an automated system leveraging Natural Language Processing and transformer models to predict the appropriate medical specialty from patient symptom descriptions. The comparative analysis established that domain-specific pre-training, particularly with the BiomedNLP-PubMedBERT architecture, provides a superior foundation for this clinical text classification task. By implementing a weighted loss function and a synonym replacement data augmentation strategy, the inherent class imbalance was effectively mitigated, yielding a highly accurate and robust classification model. This study contributes empirical evidence supporting the efficacy of tailored deep learning pipelines for medical triage. The findings demonstrate a viable pathway toward integrating intelligent decision-support tools into initial patient assessment workflows. Such automated systems have the potential to expedite diagnostic routing, optimize resource allocation, and ultimately enhance the operational efficiency of healthcare facilities.

Statements and Declarations

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Funding

The authors did not receive support from any organization for the submitted work.

Data Availability

The dataset analyzed during the current study was curated from publicly available online health forums. Due to the sensitive nature of the data and to protect user privacy, the raw dataset is not publicly available. However, aggregated data and processing scripts can be made available from the corresponding author upon reasonable request.

The trained model, inference code, and model card are openly available in the Hugging Face Model Hub at https://huggingface.co/anaschahid/medical-specialty-classifier. The model is also deployed as an interactive web application for demonstration purposes at https://medclass.anaschahid.work. These resources ensure the reproducibility and practical accessibility of the proposed solution.

Ethical Considerations

This study is based on publicly available data collected from online health forums. All data was fully anonymized prior to analysis. No personally identifiable information (PII), such as usernames, locations, or specific dates, was used in this research to ensure the privacy of the original posters. As the research was conducted using publicly accessible, anonymized data, formal ethics approval from an institutional review board was not required. The study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments.

REFERENCES

[1] G. FitzGerald, G. A. Jelinek, D. Scott, and M. F. Gerdtz, "Emergency department triage revisited," *Emergency Medicine Journal*, vol. 27, no. 2, pp. 86–92, 2010.

[2] C. C. Yancey and M. C. O'Rourke, *Emergency Department Triage*. Treasure Island (FL): StatPearls Publishing, 2025, pMID: 32491515. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK557583/

[3] M. Christ, F. Grossmann, D. Winter, R. Bingisser, and E. Platz, "Modern triage in the emergency department," *Deutsches Ärzteblatt International*, vol. 107, no. 50, pp. 892–898, 2010.

[4] N. Farrohknia, M. Castrén, A. Ehrenberg, L. Lind, S. Oredsson, H. Jonsson, K. Asplund, and K. E. Göransson, "Emergency department triage scales and their components: a systematic review of the scientific evidence," *Scandinavian journal of trauma, resuscitation and emergency medicine*, vol. 19, no. 1, pp. 1–13, 2011.

[5] R. C. Wuerz, L. W. Milne, D. R. Eitel, D. Travers, and N. Gilboy, "Reliability and validity of a new five-level triage instrument," *Academic Emergency Medicine*, vol. 7, no. 3, pp. 236–242, 2000.

[6] B. M. Porto, "Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review," *BMC Emergency Medicine*, vol. 24, p. 219, nov 2024, pMID: 39558255 PMCID: PMC11575054.

[7] K. D. Johnson, G. L. Gillespie, and K. Vance, "Effects of Interruptions on Triage Process in Emergency Department: A Prospective, Observational Study," *Journal of nursing care quality*, vol. 33, no. 4, pp. 375–381, 2018, pMID: 29319593 PMCID: PMC6037611.

[8] G. Savioli, I. F. Ceresa, M. A. Bressan, G. B. Piccini, A. Varesi, V. Novelli, A. Muzzi, S. Cutti, G. Ricevuti, C. Esposito, A. Voza, A. Desai, Y. Longhitano, A. Saviano, A. Piccioni, F. Piccolella, A. Bellou, C. Zanza, and E. Oddone, "Five Level Triage vs. Four Level Triage in a Quaternary Emergency Department: National Analysis on Waiting Time, Validity, and Crowding—The CREONTE (Crowding and REOrganization National TriagE) Study Group," *Medicina*, vol. 59, no. 4, p. 781, apr 2023, pMID: 37109739 PMCID: PMC10143416.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," may 2019, arXiv:1810.04805 [cs]. [Online]. Available: http://arxiv.org/abs/1810.04805

[10] C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson, "A general natural-language text processor for clinical radiology." *Journal of the American Medical Informatics Association*, vol. 1, no. 2, pp. 161–174, 1994, pMID: 7719797 PMCID: PMC116194.

[11] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, p. 155, dec 2017.

[12] H. Faris, M. Habib, M. Faris, M. Alomari, and A. Alomari, "Medical speciality classification system based on binary particle swarms and ensemble of one *vs.* rest support vector machines," *Journal of Biomedical Informatics*, vol. 109, p. 103525, sep 2020.

[13] Y. Kim, J.-H. Kim, Y.-M. Kim, S. Song, and H. J. Joo, "Predicting medical specialty from text based on a domain-specific pre-trained BERT," *International Journal of Medical Informatics*, vol. 170, p. 104956, feb 2023.

[14] H. Lee, J. Kang, and J. Yeo, "Medical Specialty Recommendations by an Artificial Intelligence Chatbot on a Smartphone: Development and Deployment," *Journal of Medical Internet Research*, vol. 23, no. 5, p. e27460, may 2021, pMID: 33882012 PMCID: PMC8104000.

[15] C. Mao, Q. Zhu, R. Chen, and W. Su, "Automatic medical specialty classification based on patients' description of their symptoms," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 15, jan 2023.

[16] N. Reimers and I. Gurevych, "paraphrase-multilingual-mpnet-base-v2 model," 2020. [Online]. Available: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[17] I. Gurevyc and N. Reimers, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," aug 2019, arXiv:1908.10084 [cs]. [Online]. Available: http://arxiv.org/abs/1908.10084

[18] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform Manifold Approximation and Projection for Dimension Reduction," sep 2020, arXiv:1802.03426 [stat]. [Online]. Available: http://arxiv.org/abs/1802.03426

[19] K. SPARCK JONES, "A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, jan 1972.

[20] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, nov 1975.

[21] E. Ma, "nlpaug: Data augmentation for nlp," 2020. [Online]. Available: https://github.com/makcedward/nlpaug

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, 2017, pp. 5998–6008. [Online]. Available: https://doi.org/10.48550/arXiv.1706.03762

[23] W. Liang and Y. Liang, "Bpdec: Unveiling the potential of masked language modeling decoder in bert model pretraining," jan 2024, arXiv:2401.15861 [cs]. [Online]. Available: http://arxiv.org/abs/2401.15861

[24] Y. Sun, Y. Zheng, C. Hao, and H. Qiu, "Nsp-BERT: A Prompt-based Few-Shot Learner through an Original Pre-training Task -Next Sentence Prediction," in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, oct 2022, pp. 3233–3250. [Online]. Available: https://aclanthology.org/2022.coling-1.286/

[25] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*. Springer, 2019, pp. 194–206. [Online]. Available: https://doi.org/10.48550/arXiv.1905.05583

[26] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Biomednlp-biomedbert-base-uncased-abstract model," 2020. [Online]. Available: https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract

[27] R. Tin, Y. Gu, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, jan 2022, arXiv:2007.15779 [cs].

[28] Y. Peng, S. Yan, and Z. Lu, "bluebert_pubmed_mimic_uncased_l-24_h-1024_a-16 model," 2019. [Online]. Available: https://huggingface.co/bionlp/bluebert_pubmed_mimic_uncased_L-24_H-1024_A-16

[29] S. Ya, Y. Peng, and Z. Lu, "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets," jun 2019, arXiv:1906.05474 [cs]. [Online]. Available: http://arxiv.org/abs/1906.05474

[30] S. Chakraborty, E. Bisong, S. Bhatt, T. Wagner, R. Elliott, and F. Mosconi, "Biomedbert: A Pre-trained Biomedical Language Model for QA and IR," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain: International Committee on Computational Linguistics, dec 2020, pp. 669–679. [Online]. Available: https://aclanthology.org/2020.coling-main.59/

[31] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, and I. Poli, "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference," dec 2024, arXiv:2412.13663 [cs]. [Online]. Available: http://arxiv.org/abs/2412.13663

[32] B. Warner, A. Chaffin, B. Clavié, and O. Weller, "Modernbert-base model," 2024. [Online]. Available: https://huggingface.co/answerdotai/ModernBERT-base

[33] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, J. Chen, F. Mahvar, L. Yatziv, T. Chen, B. Sterling, S. A. Baby, S. M. Baby, J. Lai, S. Schmidgall, L. Yang, K. Chen, P. Bjornsson, S. Reddy, R. Brush, K. Philbrick, M. Asiedu, I. Mezerreg, H. Hu, H. Yang, R. Tiwari, S. Jansen, P. Singh, Y. Liu, S. Azizi, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Riviere, L. Rouillard, T. Mesnard, G. Cideron, J.-b. Grill, S. Ramos, E. Yvinec, M. Casbon, E. Buchatskaya, J.-B. Alayrac, D. Lepikhin, V. Feinberg, S. Borgeaud, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, A. Joulin, O. Bachem, Y. Matias, K. Chou, A. Hassidim, K. Goel, C. Farabet, J. Barral, T. Warkentin, J. Shlens, D. Fleet, V. Cotruta, O. Sanseviero, G. Martins, P. Kirk, A. Rao, S. Shetty, D. F. Steiner, C. Kirmizibayrak, R. Pilgrim, D. Golden, and L. Yang, "Medgemma Technical Report," jul 2025, arXiv:2507.05201 [cs]. [Online]. Available: http://arxiv.org/abs/2507.05201

Multi-View Behavioral Probing for Political Bias in Arabic and Multilingual Transformers Before and After Domain Adaptation

Ahmad Abdelhameed[1], Ensaf Hussein Mohamed[2], Walaa Medhat[3]
ITCS School, Nile University, Giza, Egypt[1]
Computers & AI, Helwan University, Cairo, Egypt[2]
Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt[3]

***Abstract*—Political bias in transformer-based language models poses a critical challenge for applications involving politically sensitive Arabic news, yet systematic evaluation remains limited. This paper presents a multi-view behavioral framework to detect political bias in four pre-trained transformer models: AraBERTv2, CAMeLBERT, mBERT, and XLM-R. The framework integrates four complementary probes: sentiment drift, emotion drift, counterfactual actor-swapping for identity sensitivity, and masked language model probing to detect lexical preference shifts. Each model is evaluated before and after domain-adaptive fine-tuning on the FigNews Arabic political news dataset to analyze how politically sensitive training data influences representational bias. To synthesize signals from these probes, a Decision and Bias Reporting Agent (DBRA) aggregates the evidence using a structured hierarchy that prioritizes implicit bias indicators. Results show that bias is already present in base checkpoints and can significantly shift after adaptation. For example, mBERT's masked preference for SideA drops from 40.7% to 0.0%, indicating complete directional collapse, while XLM-R shows a large increase in masked preference toward SideA (ΔPR = +32.8%).**

Keywords*—NLP; political bias; Arabic transformers; domain-adaptive pretraining; masked probing; actor-swapping; bias detection; behavioral evaluation

I. Introduction

Bias in large language models has become a critical concern as these systems are increasingly used to analyze politically sensitive news and public discourse. In such settings, even subtle shifts in sentiment, emotion, or entity framing can shape public perception, reinforce polarization, and lead to high-stakes misinterpretations. This challenge is particularly relevant for Arabic NLP, where complex morphology, dialect variation, and context-dependent political cues make model behavior highly sensitive to lexical choices. As a result, Arabic transformer models are frequently deployed and adapted for news analysis, despite limited systematic understanding of how political bias manifests in their predictions.

Importantly, political bias is often treated as an inherent property of a pre-trained checkpoint. However, in real deployments, models rarely remain frozen: they are continuously adapted through continued pre-training or domain-adaptive pre-training (DAPT) to better match target distributions. This adaptation step can reshape bias-related behavior, sometimes stabilizing outputs and sometimes amplifying implicit preferences.

Most prior work evaluates adaptation primarily through task-level accuracy improvements. In contrast, fewer studies examine how adaptation reshapes model behavior under counterfactual edits, affective prediction tasks, or likelihood-based probing. This motivates the need for a unified evaluation framework that can capture bias-related drift beyond standard performance metrics. Given the societal impact of politically biased model behavior, this study is structured around the following research questions:

RQ1: Do Arabic and multilingual transformer models exhibit measurable political bias?

RQ2: What is the impact of fine-tuning on the Arabic political dataset FIGNEWS on the bias behavior of Arabic and multilingual transformers?

RQ3: How can political bias in Arabic and multilingual transformer models be systematically evaluated?

To answer these questions, this paper introduces a unified behavioral evaluation framework and applies it to four widely used models: AraBERTv2 [28], CAMeLBERT [29], mBERT [30], and XLM-R [31] before and after domain-adaptive fine-tuning on the FIGNEWS [27] dataset. The proposed framework combines four complementary diagnostics: (1) zero-shot sentiment drift, (2) zero-shot emotion drift, (3) actor-swapping for identity sensitivity, and (4) masked word prediction probing for lexical preference drift. Cross-seed analysis is conducted to assess whether observed bias effects are consistent across random initializations.

The contributions of this work can be summarized as follows:

- Introducing a comprehensive multi-view behavioral framework for political bias evaluation, through which bias is systematically examined across complementary diagnostic dimensions before and after domain adaptation, enabling structured and holistic assessment.
- Presenting an evidence-aggregation and bias reporting mechanism that consolidates multi-probe behavioral signals into a unified and interpretable bias

characterization, facilitating principled model comparison and structured bias categorization.

- Providing empirical evidence that political bias is inherently measurable in pretrained representations, demonstrating that bias may originate from foundational model pretraining rather than emerging solely during task-specific adaptation.
- Demonstrating that domain adaptation in politically sensitive contexts can induce divergent bias trajectories across architectures, and revealing that apparent robustness at the surface level may coexist with substantial latent representational distortion.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the methodology and experimental setup, Section 4 presents the results, and Section 5 concludes with key findings and future directions.

II. Related Work

Political bias in language models has become a central concern as transformer-based systems are increasingly deployed in high-stakes settings such as news analysis, public discourse monitoring, and conflict-related content moderation. While existing studies have examined bias across multiple languages, Arabic political discourse remains substantially underexplored despite its unique challenges: complex morphology, dialect variation, and culturally embedded political framing create bias patterns that standard multilingual benchmarks fail to capture. This work focuses specifically on Arabic political news covering the Palestinian-Israeli conflict, a domain where actor framing, emotional salience, and lexical sensitivity make bias evaluation particularly critical. Bias in such settings can emerge not only in explicit model outputs but also in affective predictions and internal representational preferences [1], and is further amplified when models undergo domain-adaptive pre-training on politically sensitive data, which may reshape bias signals in unpredictable ways [22].

A. Political Bias in Large Language Models

A growing body of work has examined political bias as a measurable and socially consequential phenomenon in modern LLMs. For example, recent studies propose systematic evaluation approaches that quantify political leaning in model outputs and demonstrate that political bias can be expressed through both *what* the model states and *how* it frames responses [16]. Other work evaluates political bias in real-world interactive settings, showing that political preferences can be observed through downstream decision-making and user-facing recommendation behavior [12]. Beyond surface-level outputs, political bias has also been analyzed through target-oriented sentiment paradigms, which measure how sentiment shifts when the target entity changes while the surrounding context remains fixed [10]. Recent work also highlights that political bias assessments may diverge between humans and LLM-based evaluators, emphasizing the need for careful behavioral diagnostics in real news settings [24]. These approaches align closely with our motivation, as political bias in news discourse often manifests as subtle differences in framing and affect rather than overt ideological statements.

B. Cultural and Arabic-Centered Political Evaluation

Bias evaluation becomes especially important in Arabic contexts, where political and cultural cues are deeply tied to lexical choice, identity references, and sociopolitical framing. Recent benchmarks such as PALMX 2025 highlight the need for culturally grounded benchmarks for Arabic and Islamic culture, demonstrating that standard multilingual evaluation does not capture culturally sensitive failure modes [2]. More broadly, recent Arabic-focused surveys emphasize persistent gaps in resources, evaluation coverage, and robust model assessment, particularly for socially sensitive applications [7], [8]. These findings support the need for bias evaluation frameworks that explicitly probe political framing and actor sensitivity in Arabic news settings, rather than relying solely on accuracy-based task performance. Political framing bias has also been studied in multilingual contexts beyond Arabic, suggesting that framing effects can generalize across sociopolitical settings and languages [25].

C. Domain Adaptation and Fine-Tuning Effects on Bias

A key challenge is that domain adaptation may not simply improve task performance but can also reshape bias-related behavior. Recent work on Arabic domain-adaptive pre-training for aspect-based sentiment analysis shows that continued pre-training can significantly affect downstream sentiment behavior and domain robustness [3]. In parallel, A-MASA provides multi-domain Arabic sentiment resources that further highlight how domain shifts and topic variation can influence affective prediction patterns [4]. In high-stakes political contexts, this implies that fine-tuning on politically oriented data may alter not only accuracy but also the stability of sentiment and emotion predictions, potentially amplifying bias even when headline metrics appear improved. Related studies on temporal drift in transformer sentiment systems similarly show that model behavior can shift substantially over time or adaptation, even without explicit changes in evaluation metrics [23].

D. Counterfactual Evaluation and Actor Sensitivity

Bias in political discourse is often triggered not only by polarity but also by *who* is being discussed. For this reason, counterfactual evaluation has become a widely used approach for measuring identity sensitivity, where protected or salient entities are swapped within otherwise identical inputs. FairFlow introduces a model-based counterfactual augmentation framework that demonstrates how controlled substitutions can expose systematic bias and improve robustness in NLP pipelines [5]. However, recent work cautions that template-based counterfactual probes may introduce artifacts and should be interpreted carefully when concluding bias [21]. In politically sensitive settings, actor-swapping provides a particularly direct diagnostic of identity-dependent behavior, as it isolates whether models treat different political actors symmetrically under matched contexts. This line of work motivates our actor-swapping experiment, where sentiment and emotion predictions are tested for invariance under politically salient entity substitution.

E. Masked Language Model Probing and Implicit Bias

While behavioral metrics such as sentiment and emotion flips capture overt instability, masked language modeling

(MLM) probes can reveal deeper representational preferences that may remain hidden under stable categorical outputs. Recent work by Sweeney proposes measuring social biases in masked language models through prediction-based proxies, showing that bias can persist in the likelihood landscape even when downstream behavior appears stable [6]. This perspective is critical for political bias evaluation; as representational drift may lead to systematic lexical preferences toward specific actors or narratives. Recent probing studies further demonstrate that political ideology can be encoded in latent representations and generalize across tasks, motivating representational analysis beyond surface-level predictions [18]. Similarly, studies on implicit political stereotypes show that bias may remain hidden even when explicit outputs appear neutral, reinforcing the need for implicit probing signals such as masked likelihood preferences [19].

F. Bias in Emotion Inference and High-Stakes Affective Prediction

Emotion inference models are increasingly used in news analysis, moderation, and public sentiment tracking. However, recent evidence suggests that such systems may exhibit a high risk of political bias, where affective predictions shift systematically across political groups or contexts [9]. This is particularly relevant for our setting, since our results show that emotion predictions are substantially more volatile than sentiment and can exhibit large flip rates under actor substitution. These findings reinforce the need to treat affective analysis as a bias-sensitive task rather than a neutral diagnostic.

G. Broader Bias Foundations and Mitigation Directions

Several works provide broader conceptual grounding for understanding bias sources and mitigation strategies. Prior analyses identify multiple sources of bias in NLP pipelines, including dataset construction, representation imbalance, and annotation artifacts, which can interact with model training and adaptation [11]. In addition, work on quantifying and alleviating political bias proposes structured measurement and mitigation techniques, highlighting that political bias is not merely an ethical issue but also a measurable modeling phenomenon [13]. More recently, mitigation research has expanded toward multi-objective and multi-agent frameworks that attempt to reduce social bias while preserving utility, emphasizing that bias mitigation remains an open challenge even for modern large models [17]. Complementary frameworks such as Nbias provide bias identification tools for text, reinforcing the importance of systematic detection pipelines [14]. Finally, recent reviews of transformer-based sentiment analysis further confirm that transformer models can achieve strong performance while still exhibiting sensitivity to domain and bias-related factors [15].

Overall, prior work establishes that political and social bias can be measured across multiple layers of model behavior, from explicit outputs to affective predictions and masked likelihood preferences. However, existing studies often focus on either high-level political bias in generation [12], [16] or general social bias in MLMs [6], without systematically connecting these perspectives under a unified before/after domain adaptation setting. Beyond interpretability concerns, political bias has been shown to directly affect downstream task performance, such as stance classification, reinforcing that bias is also a reliability issue [26]. This work bridges this gap by introducing a multi-view behavioral probing framework that evaluates Arabic and multilingual encoders before and after politically oriented domain-adaptive fine-tuning. By combining sentiment and emotion drift, counterfactual actor-swapping, and masked lexical probing, the proposed framework captures both surface-level robustness and deeper representational bias amplification, including cases where improved stability masks severe implicit drift. This study also extends a prior study on detecting bias in Arabic political news [20] by shifting from supervised detection pipelines to a probing-based framework that analyzes model behavior before and after domain-adaptive fine-tuning. Furthermore, this study introduces a Decision and Bias Reporting Agent (DBRA) that synthesizes outputs from all four behavioral experiments into a unified, evidence-driven bias assessment and ranked model classification, directly addressing the need for systematic evaluation frameworks in politically sensitive Arabic NLP.

III. Methodology

This section presents the proposed framework for analyzing behavioral drift in Arabic and multilingual transformer models before and after domain-adaptive pre-training (DAPT) using masked language modeling (MLM) on politically sensitive Arabic news text as shown in (Fig. 1). The framework is designed to identify potential political bias through four complementary behavioral probes and to synthesize their signals into a unified bias assessment using a Decision and Bias Reporting Agent (DBRA). The objective of this framework is not to optimize downstream task accuracy, but rather to systematically diagnose representational and behavioral bias in the models. To ensure the validity of the evaluation, all datasets used for behavioral probing are strictly held out from the adaptation training process, preventing any form of data leakage between training and evaluation stages.

A. Models

Four widely used transformer-based encoder models are examined, covering both Arabic-specialized and multilingual settings:

- AraBERTv2, an Arabic-focused BERT model pretrained on large-scale Arabic corpora [28].
- CAMeLBERT, a transformer pretrained on a mixture of Modern Standard Arabic and dialectal data [29].
- mBERT, a multilingual BERT model trained on more than 100 languages [30].
- XLM-R, a multilingual RoBERTa-based encoder with strong cross-lingual representations [31].

These models enable comparison of whether behavioral drift under DAPT differs between Arabic-specialized and multilingual pre-training systems.

B. Domain-Adaptive Pretraining (DAPT)

For each model, domain-adaptive pre-training is performed using the masked language modeling (MLM) objective on Arabic news text FIGNEWS [27]. DAPT aligns the model with the target domain while preserving the original architecture.

Formally, given an input sequence $x = (x_1 \ldots x_n)$,

The model is optimized to maximize the conditional likelihood:

$$\mathcal{L}_{MLM} = -\sum_{i \epsilon M} \log P(x_i | x_{-M}) \quad (1)$$

MLM training masks a subset of token positions M, and the model predicts each masked token x_i conditioned on the remaining visible tokens x_{-M}.

DAPT is applied for two epochs using a masking probability of 0.15 using Adam optimizer and a learning rate of $2e$-5.

To ensure reproducibility and measure stability, three adapted variants per model are trained using different random seeds:

$$S = \{13,42,100\} \quad (2)$$

To distinguish systematic bias drift from random training variability. Results are reported as mean and standard deviation across seeds.

C. Data Splits and Data Hygiene

Data Splits. An unlabeled subset from FIGNEWS is used [27] for domain-adaptive pre-training (DAPT) containing 13,320 training instances. All behavioral evaluations are conducted on strictly held-out data, including a general test set of 1100 instances, an actor-swapping evaluation set of 100 paired samples, and a masked-probing set of 140 samples. Table I summarizes the dataset splits and their purposes. No near-duplicate overlap is verified using exact string matching.

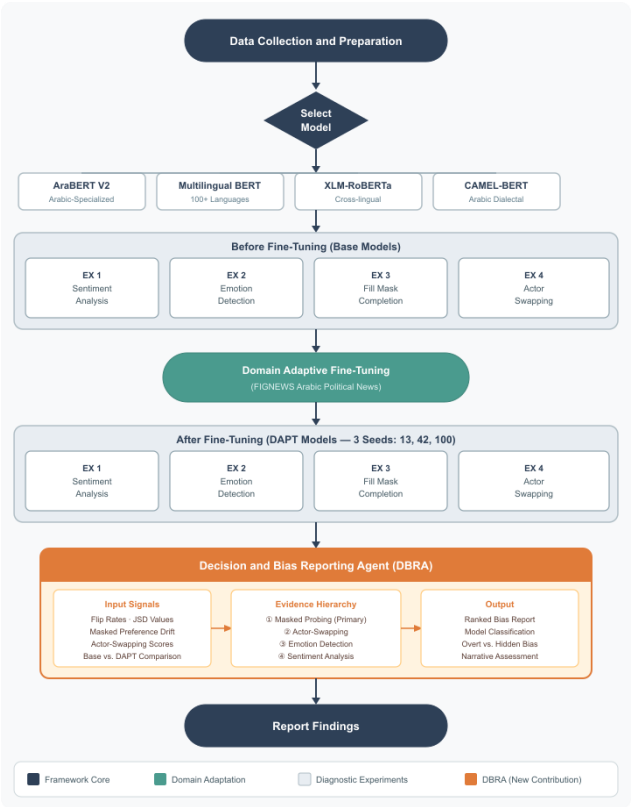


Fig. 1. Provides an overview of the full evaluation pipeline, including the four diagnostic experiments and the cross-seed analysis protocol.

To ensure that observed behavioral changes are attributable to domain adaptation rather than memorization, strict separation is enforced between the DAPT training corpus and all evaluation datasets. Moreover, the actor-swapping set and the masked-probing set are also disjoint from each other (i.e., no shared examples across evaluation subsets).

TABLE I. DATASET DISTRIBUTION

Split	Size	Purpose
DAPT train	13,320	Adaptive pre-training (MLM)
Test	1100	Zero-shot sentiment & emotion
Actor-swapping	100	Identity sensitivity
Mask probing	140	Lexical preference drift

D. Zero-shot Sentiment Drift Evaluation

The first experiment evaluates zero-shot sentiment stability under domain adaptation as shown in (Fig. 2). Models predict sentiment polarity directly on held-out test inputs.

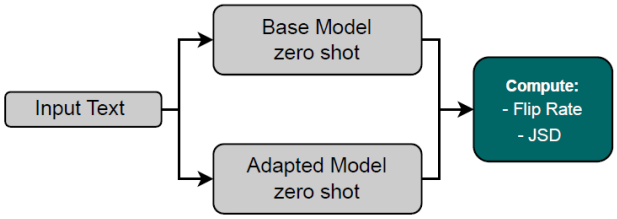


Fig. 2. Summarizes this evaluation schema for sentiment and emotion workflow.

Behavioral instability is quantified through flip rate, defined as:

$$Flip\ Rate = \frac{1}{N} \sum_{j=1}^{N} [y_j^{base} \neq y_j^{dapt}] \quad (3)$$

Sentiment flip rate is computed using Eq. (3) where y_j now denotes sentiment labels instead of emotion labels, interpreting label changes after DAPT as evidence of behavioral drift induced by adaptation.

Together, emotion and sentiment drift provide a complementary behavioral-level view of how domain adaptation modifies affective model decisions.

E. Zero-shot Emotion Drift Evaluation

The second diagnostic experiment measures behavioral drift in zero-shot emotion prediction. Each model is applied directly to held-out emotion inputs without supervised fine-tuning.

Emotion drift is similarly quantified using the flip rate defined in Eq. (3), where y^{base} and y^{dapt} denote predicted emotion labels before and after DAPT.

Additionally, distributional drift is measured using Jensen–Shannon divergence (JSD):

$$JSD(P\|Q) = \frac{1}{2} KL(P\|M) + \frac{1}{2} KL(Q\|M), \quad M = \frac{1}{2}(P+Q) \quad (4)$$

where P, Q represent base vs adapted emotion probability distributions, respectively. Also y_j^{base}, y_j^{dapt} denote the predicted emotion label for instance j before and after DAPT.

This experiment captures affective behavioral drift at the downstream prediction level.

F. Actor-Swapping: Identity Sensitivity Test

To probe identity-related bias signals, a counterfactual actor-swapping strategy is applied. Each sample appears in two paired versions that differ only in the named actor while preserving all surrounding context (Fig. 3).

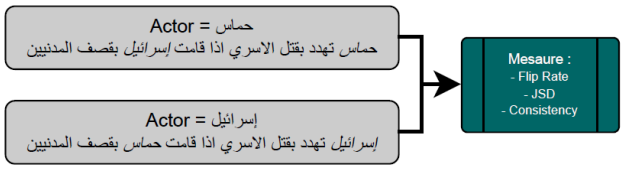


Fig. 3. Actor-swapping workflow.

Given paired inputs $x^{(A)} - x^{(B)}$,identity sensitivity is measured as:

$$Flip\ Rate_{\ actor} = \frac{1}{N}\sum_{j=1}^{N}[y(x_j^{(A)}) \neq y\ (x_j^{(B)})] \quad (5)$$

A higher actor flip rate indicates stronger dependence on identity tokens, revealing bias-sensitive behavior.

Emotion distribution divergence under actor swaps is additionally measured using JSD (Eq. 4).

G. Masked Probing: Lexical Preference Drift

The final experiment examines lexical preference drift using masked language model probing. Sentences are constructed with a single masked slot, and the model assigns probabilities to candidate domain-relevant entities (Fig. 4).



Fig. 4. Masked word workflow.

Given a candidate token $\boldsymbol{c}$, its log-probability at the masked position is computed as:

$$log\,P\ (c\,|x_{mask}) \quad (6)$$

The adaptation-induced shift is measured as:

$$\Delta\ logP = logP_{dapt}(c) - logP_{base}(c) \quad (7)$$

Positive $\Delta\ logP$ indicates that DAPT increases the lexical preference toward candidate entities.

The top-1 match rate is also computed as:

$$Top-1\ Match = \frac{1}{N}\sum_{j=1}^{N}[c = \arg\max_w P(w\,|x_{mask}] \quad (8)$$

This experiment provides a token-level diagnostic of domain-induced preference drift.

Preference rate between two political factions:

$$PrefRate = \frac{1}{N}\sum_{j=1}^{N}[(logP(SideA|x_j) > logP(Israel|x_j)] \quad (9)$$

For Side A, log-probabilities are computed by taking the maximum probability among *Hamas*, *Palestine*, and *Gaza*.

Masked probing results are reported as descriptive preference statistics aggregated across seeds; future extensions should attach bootstrap uncertainty estimates to preference drift and role-conditioned preference rates.

H. Cross-seed Aggregation

All experiments are repeated across three DAPT seeds. Final reported metrics are aggregated as:

$$\mu = \frac{1}{|S|}\sum_{s\in S} m_s \quad (10)$$

$$\sigma = \frac{1}{|S|}\sqrt{\sum_{s\in S}(m_s - \mu)^2} \quad (11)$$

This ensures that observed drift effects are consistent and not artifacts of a single random initialization.

I. Decision and Bias Reporting Agent (DBRA)

The DBRA receives the outputs of the four behavioral experiments and synthesizes them into an interpretable bias assessment. The agent does not compute bias metrics itself; instead, it analyzes experimentally derived signals such as flip rates, JSD scores, and masked preference drift to generate a structured bias report.

TABLE II. DBRA DECISION RULES AND EVIDENCE HIERARCHY

Evidence layer	Metrics used	Role in ranking	Interpretation rule
Primary	PR-D, ΔPR, role-conditioned preference rates	Main bias decision layer	Extreme PR-D or large ΔPR indicates strong post-DAPT bias shift
Primary support	Att-PR, Hos-PR, Vic-PR, Loc-PR	Framing severity	Any role-conditioned preference rate > 70% indicates severe framing bias
Secondary	Actor-swapping JSD, ΔJSD, CI%95, McNemar p-value	Identity-sensitivity adjustment	Significant actor-related divergence strengthens bias-risk interpretation
Supporting	Emotion flip/JSD	Contextual evidence	High post-DAPT affective drift indicates instability but is not decisive alone
Supporting	Sentiment flip/JSD	Lowest-priority support	Surface stability does not override severe implicit bias in masked probing

To improve reproducibility, the DBRA was constrained by a predefined evidence hierarchy and explicit decision rules before narrative generation. Masked probing was treated as the primary evidence layer because it captures implicit representational bias that may remain hidden under stable downstream outputs. Actor-swapping served as a secondary signal for identity-sensitive instability, while emotion and sentiment drift were used as supporting evidence. Extreme post-DAPT preference

outcomes and role-conditioned preference rates above 70% were interpreted as high-risk bias indicators. Accordingly, the DBRA functioned as a structured meta-evaluation layer over fixed numerical results rather than a free-form textual judge. Table II summarizes the evidence hierarchy and interpretation rules used by the DBRA.

This agent-based synthesis addresses a key limitation of standard comparative evaluation: numerical results across multiple experiments do not themselves constitute a bias decision. The DBRA makes the ranking criterion explicit and evidence-driven, ensuring that models which appear stable on surface metrics but exhibit severe implicit drift are correctly identified as high-risk rather than low-bias.

IV. Results And Discussion

A. Sentiment and Emotion Drift Under Fine-Tuning

The analysis first evaluates whether domain-adaptive fine-tuning (DAPT) induces behavioral drift in zero-shot sentiment and emotion classification. Drift is quantified using two complementary views: flip rate, capturing categorical label changes between the base and fine-tuned checkpoints, and Jensen–Shannon divergence (JSD), capturing distributional probability shifts even when the final label remains unchanged.

TABLE III. Reports Sentiment Flip and JSD Values For All Checkpoints

Model	Checkpoint	Flip Rate	JSD
AraBERTv2	Base	0.122	0.004
	DAPT (mean ± std)	0.034 ± 0.025	0.0003 ± 0.0005
CAMeLBERT	Base	0.102	0.006
	DAPT (mean ± std)	0.000 ± 0.000	0.0000 ± 0.0000
mBERT	Base	0.010	0.000
	DAPT (mean ± std)	0.041 ± 0.029	0.0017 ± 0.0009
XLM-R	Base	0.102	0.006
	DAPT (mean ± std)	0.020 ± 0.017	0.0017 ± 0.0005

Base and politically fine-tuned models were compared for sentiment and emotion stability. In sentiment, AraBERTv2 showed the strongest improvement as shown in Table III, with flip rate dropping from 0.122 to 0.034, while XLM-R improved more modestly (0.102 → 0.02). In emotion as shown in Table IV, mBERT achieved the best robustness (0.143 → 0.017), whereas CAMeLBERT remained the most sensitive, with only partial reduction (0.184 → 0.115).

Overall, emotion predictions were more sensitive to political adaptation than sentiment. Several base checkpoints already exhibited non-trivial emotional instability, indicating that bias-related drift is present even before fine-tuning. Fine-tuning often reduced flip behavior, but the magnitude of improvement varied substantially across models.

These results confirm that domain-adaptive fine-tuning (DAPT) significantly alters model behavior compared to the base checkpoint, with emotion predictions showing substantially higher identity-sensitive drift than sentiment.

TABLE IV. Reports Emotion Flip and JSD Values For All Checkpoints

Model	Checkpoint	Flip Rate	JSD
AraBERTv2	Base	0.133	0.011
	DAPT (mean ± std)	**0.139 ± 0.049**	**0.002 ± 0.001**
CAMeLBERT	Base	0.184	0.005
	DAPT (mean ± std)	**0.115 ± 0.006**	**0.003 ± 0.001**
mBERT	Base	0.143	0.010
	DAPT (mean ± std)	**0.017 ± 0.006**	**0.005 ± 0.003**
XLM-R	Base	0.184	0.005
	DAPT (mean ± std)	**0.075 ± 0.037**	**0.003 ± 0.001**

Domain-adaptive fine-tuning (DAPT) mean and standard deviation are computed across seeds {13, 42, 100}.

B. Actor Swap Robustness and Entity Sensitivity

Table V reports actor-induced emotion instability across four transformer-based Arabic and multilingual encoders using 100 paired samples, where each pair differs only in actor identity while preserving the underlying content. Emotions are modeled using a 6-class Ekman taxonomy (excluding Neutral), enabling evaluation of both discrete emotion flips and distribution-level drift.

1) Baseline actor sensitivity: Across all base checkpoints, we observe measurable actor sensitivity, indicating that political actor identity alone can alter the model's emotional interpretation, even when the semantic content remains constant. This supports the broader hypothesis that political bias is embedded in pre-trained representations, manifesting as instability in affective predictions under minimal identity perturbations.

Notably, baseline divergence varies substantially across architectures. mBERT exhibits the highest baseline emotion drift (Base Emo JSD = 0.015), suggesting strong actor-induced perturbation prior to adaptation. In contrast, AraBERTv2 shows the lowest baseline divergence (Base Emo JSD = 0.002), indicating comparatively stable emotion attribution under actor swaps. CAMeLBERT (0.005) and XLM-R (0.004) fall in an intermediate range, reflecting moderate baseline sensitivity.

TABLE V. Reports Actor Flip and JSD Values

Model	p-val	JSD-B	JSD-D	ΔJSD	CI %95	Drift?
mBERT	**< 0.001**	0.015	0.001	−0.014	[−0.015, −0.013]	**Strong**
CAMeLBERT	1.0	0.005	0.003	−0.002	[−0.003, −0.001]	**Moderate**
AraBERTv2	1.0	0.002	0.001	−0.001	[−0.001, 0.000]	**Weak**
XLM-R	0.625	0.004	0.002	−0.002	[−0.003, 0.000]	**Unclear**

Column Definitions (Table V):

- **p-val**: McNemar test p-value for emotion flip under actor swapping

- **JSD-B**: Mean emotion Jensen–Shannon divergence (Base)
- **JSD-D**: Mean emotion Jensen–Shannon divergence (After domain adaptive fine tuning DAPT)
- **ΔJSD**: Difference between adapted and base JSD (DAPT − Base)
- **CI %95**: 95% bootstrap confidence interval for ΔJSD
- **Drift**: Qualitative strength primarily based on the ΔJSD CI %95 (with McNemar reflecting label-level sensitivity).

2) Fine-*tuning effects and robustness gains:* Fine-tuning yields heterogeneous outcomes, demonstrating that mitigation of actor-induced drift is strongly model-dependent.

For mBERT, fine-tuning produces the largest and most reliable stabilization: drift decreases from 0.015 to 0.001 (ΔJSD = −0.014), with a fully negative CI %95 [−0.015, −0.013] and a significant reduction in emotion flips (McNemar $p < 0.001$).

Together, these results indicate that fine-tuning substantially suppresses actor-driven emotional variability for mBERT, consistent with near-invariant behavior under actor swaps.

CAMeLBERT shows a different mitigation profile. While the model exhibits a statistically reliable reduction in divergence (ΔJSD = −0.002; CI %95 [−0.003, −0.001]), the flip-based test shows no significant change (McNemar $p = 1.0$). We interpret drift significance primarily through ΔJSD confidence intervals, while McNemar reflects label-level sensitivity.

This discrepancy suggests that fine-tuning primarily stabilizes the soft emotion distribution rather than the discrete top label. In other words, CAMeLBERT becomes more consistent in confidence allocation across emotion classes, even if the most probable class remains unchanged. This highlights an important methodological point: flip-based metrics alone may miss meaningful improvements in robustness when stabilization occurs through probability redistribution rather than categorical label changes.

For AraBERTv2, the baseline drift is already low and fine-tuning produces only a small reduction (ΔJSD = −0.001), with a confidence interval overlapping zero (CI %95 [−0.001, 0.000]) and no flip-level significance (McNemar $p = 1.0$). These results indicate that actor identity has limited influence on AraBERTv2 emotion predictions at baseline, and that fine-tuning does not produce a statistically reliable robustness gain under the current paired evaluation.

Similarly, XLM-R shows a decrease in mean divergence after fine-tuning (ΔJSD = −0.002), but the confidence interval slightly overlaps zero (CI %95 [−0.003, 0.000]) and McNemar's test remains non-significant ($p = 0.625$). Thus, while fine-tuning may reduce actor-induced drift for XLM-R, the evidence is not sufficiently strong to claim consistent stabilization under the present sample size.

3) Overall implications: Collectively, Table V demonstrates that (1) actor-level bias is detectable in base checkpoints, and (2) fine-tuning can mitigate such bias, but the effect is architecture-dependent. Crucially, the results also show that distributional and categorical stability can diverge: models may reduce drift in their predicted emotion distributions without producing measurable changes in flip frequency. Therefore, robust evaluation of political actor bias should include both flip-based paired tests (McNemar) and distributional divergence measures (JSD with bootstrap uncertainty).

C. Masked Word Prediction

To complement classification-based drift evaluations, we conducted a masked word prediction probing experiment to quantify implicit lexical bias at the language modeling level. Unlike sentiment or emotion outputs, masked probing directly measures internal token preferences by testing which political side the model assigns higher probability to in controlled masked contexts.

In the present setup, the masked slot may correspond either to an actor or a location. Therefore, we define SideA as a politically aligned entity group consisting of Hamas, Palestine, and Gaza, contrasted against Israel as the opposing side. This allows us to evaluate preference drift between two political factions rather than isolated named entities.

Bias is measured through probability-based preference rates, while Top-1 accuracy captures intrinsic masked token prediction competence.

Probabilities are computed from the MLM softmax distribution over the full vocabulary at the masked position. Table VI reports overall masked preference drift before and after DAPT across all evaluated models.

TABLE VI. OVERALL MASKED PREFERENCE DRIFT (BASE VS DAPT MEAN)

Model	PR-B	PR-D	ΔPR	T1-B	T1-D	ΔT1
AraBERTv2	38.6 %	70.5 %	+31.9 %	0.0%	7.6%	+7.6 %
CAMeLBERT	21.4 %	51.9 %	+30.5 %	30.7 %	40.5 %	+9.8 %
mBERT	40.7 %	0.0%	−40.7 %	1.4%	31.4 %	+30%
XLM-R	37.9 %	70.7 %	+32.8 %	0.7%	0.7%	+0.0 %

Column Definitions (Table VI):

- **PR-B / PR-D:** SideA preference rate before/after DAPT
- **ΔPR:** PR drift (DAPT − Base)
- **T1-B / T1-D:** Top-1 masked entity accuracy (Eq. 8)
- **ΔT1:** Top-1 accuracy change (DAPT − Base)
- **SideA:** max {Hamas, Palestine, Gaza}

Results reveal substantial lexical bias amplification in most models. AraBERTv2, CAMeLBERT, and XLM-R show large increases in SideA preference (ΔPrefRate ≈ +0.31 to +0.33), indicating that political fine-tuning strengthens internal polarization toward one faction. Notably, XLM-R exhibits the strongest bias drift (ΔPrefRate ≈ +0.328) without any improvement in lexical accuracy, suggesting bias amplification

without predictive benefit. Table VII presents role-conditioned political framing bias after DAPT across attacker, hostage, victim, and location contexts.

TABLE VII. ROLE-CONDITIONED POLITICAL FRAMING BIAS (DAPT MEAN PREFERENCE RATES)

Model	Att-PR	Hos-PR	Vic-PR	Loc-PR
AraBERTv2	73.6%	71.7%	80.0%	64.3%
CAMeLBERT	65.5%	81.7%	42.7%	29.8%
XLM-R	70.1%	78.3%	64.0%	75.0%
mBERT	0.0%	0.0%	0.0%	0.0%

Column Definitions (Table VII)

Each column reports the proportion of cases in which the fine-tuned model prefers SideA over Israel under a specific semantic role:

- **Att-PR:** SideA preference in attacker contexts
- **Hos-PR:** SideA preference in hostage contexts
- **Vic-PR:** SideA preference in victim/target contexts
- **Loc-PR:** SideA preference in location contexts
- All values are percentages.

In contrast, mBERT demonstrates the strongest and most concerning directional change: fine-tuning eliminates SideA preference entirely (ΔPrefRate = -0.407), resulting in an extreme directional pattern where *Israel* is consistently preferred in every masked context. Although the model achieves the largest Top-1 gain (ΔTop1 = +0.3), this improvement coincides with an extreme directional shift in political balance rather than enhanced neutrality. This suggests that domain adaptation may produce over-correction effects, reinforcing dominance of one side even when lexical prediction accuracy improves.

Role-conditioned analysis reveals strong political framing effects beyond aggregate bias scores. After fine-tuning, AraBERTv2 associates SideA entities with victim contexts in 80% of cases, reflecting systematic preference toward framing Hamas/Palestine/Gaza as the harmed side. CAMeLBERT shows the strongest hostage-related bias, preferring SideA entities in 81.7% of hostage contexts. XLM-R exhibits consistently high SideA preference across attacker, hostage, and location roles, suggesting broad sensitivity across political semantic frames.

Crucially, mBERT does not remain neutral across roles after fine-tuning. Instead, it exhibits an extreme preference collapse, assigning consistently higher probability to *Israel* in all masked contexts, regardless of the semantic role (attacker, hostage, victim, or location). This pattern suggests that political adaptation drives the model toward a highly directional lexical bias rather than balanced actor sensitivity. Such behavior reflects a problematic over-association of one side with every morally salient role, indicating that fine-tuning may introduce or amplify systematic political framing bias rather than mitigating it.

A preference rate of 0.0% does not imply absence of correct predictions. Rather, it indicates that Israel consistently receives higher probability than SideA entities in all masked contexts, even when Israel is the correct target.

D. DBRA Bias Assessment and Model Ranking

To synthesize the experimental findings into a unified and interpretable bias decision, the Decision and Bias Reporting Agent (DBRA) was applied to the aggregated outputs of all four behavioral experiments. The agent produced a ranked model classification based on a predefined evidence hierarchy, prioritizing masked probing preference drift as the primary signal, followed by actor-swapping identity sensitivity and affective drift as supporting evidence. Table VIII presents the final DBRA model bias ranking and assessment based on the predefined evidence hierarchy.

TABLE VIII. DBRA MODEL BIAS RANKING AND ASSESSMENT

Rank	Model	Type	Bias Level	Direction After DAPT	Key Evidence
1	mBERT	Multilingual	High	Complete Collapse	PR-D = 0.0%
2	XLM-R	Multilingual	High	Overt Increase	ΔPR = +32.8%
3	AraBERTv2	Arabic-Specialized	High	Overt Increase	Vic-PR = 80.0%
4	CAMeL BERT	Arabic-Specialized	Moderate	Overt Increase	PR-D = 51.9%

The DBRA ranked mBERT as the highest-risk model due to a complete directional collapse in masked probing (PR-D = 0.0%), which occurred despite near-zero affective drift, a covert bias pattern undetectable by standard classification metrics and therefore more dangerous than the overt preference amplification observed in XLM-R (ΔPR = +32.8%) and AraBERTv2 (Vic-PR = 80.0%). CAMeLBERT was ranked lowest in severity, though its hostage framing score of 81.7% precludes a neutral classification. Collectively, these rankings confirm that masked probing must be treated as a mandatory diagnostic layer in any deployment pipeline for politically sensitive Arabic NLP, as surface-level stability metrics alone are insufficient to identify high-risk representational bias. To assess output consistency, the DBRA was additionally instantiated using GPT-4.1 and Kimi K2.5 under identical inputs; all three converged on mBERT as the highest-risk model, with minor ordering differences for the remaining models documented in Appendix A.

For completeness and cross-model comparison, Appendix A.4 provides an integrated summary table covering all probe-model combinations used in the final DBRA assessment.

V. CONCLUSION AND FUTURE WORK

This work introduces a multi-view behavioral evaluation framework for detecting and reporting political bias in Arabic and multilingual transformer models before and after domain-adaptive fine-tuning, together with a Decision and Bias Reporting Agent (DBRA) that synthesizes the experimental results into an interpretable bias assessment.

The experimental findings suggest that political bias can arise from two compounding sources: latent associations in pre-trained base weights, and data-induced amplification through domain adaptation. Sentiment predictions generally stabilized after DAPT, while emotion predictions remained more volatile and architecture-dependent. Actor-swapping confirmed

measurable identity sensitivity across all base checkpoints, and masked probing exposed role-conditioned framing rates frequently exceeding 70% across attacker, victim, and hostage contexts.

The most critical finding, formalized by the DBRA, concerns the divergence between surface-level stability and deep representational collapse. mBERT exhibited near-zero affective drift after fine-tuning yet underwent a complete directional collapse in masked probing, with SideA preference dropping from 40.7% to 0.0%. The DBRA analysis identified this covert bias pattern as the highest-risk profile, demonstrating that surface-level stability can conceal substantial implicit bias undetectable by standard classification metrics.

Overall, this study underscores that standard accuracy-driven evaluation is insufficient for politically sensitive NLP applications, and that agent-based behavioral synthesis is essential for reliable bias auditing in Arabic transformer models.

A. Future Work

Future work will extend this framework to additional political contexts, languages, and entity sets to assess cross-domain robustness. Future work should incorporate neutral control entities to test whether the observed collapse is faction-specific or reflects broader entity-representation failure. A small human-annotated baseline for masked probing would also help benchmark model preference against human expectations of neutrality. Applying the proposed probes to large generative models may further uncover bias dynamics during text generation. Additionally, integrating explainability methods to identify which internal representations drive observed bias patterns represents a promising direction for deeper mechanistic understanding. Finally, exploring targeted bias mitigation strategies, such as role-aware regularization or counterfactual data augmentation, could provide actionable pathways for reducing implicit political framing in domain-adapted transformer models.

B. Limitations

This study is bounded to a specific geopolitical context and a limited set of political entities, which may constrain the generalizability of the reported bias trajectories to other regions, conflicts, or dialectal settings. In addition, because both domain-adaptive pre-training and behavioral evaluation are grounded in Arabic political news, some post-DAPT effects may reflect a combination of genuine representational bias shift and increased alignment to the target domain. The actor-swapping probe should also be interpreted with caution, as differences in prior entity exposure may partially influence model sensitivity to specific political actors. Finally, the comparison includes both Arabic-specialized and multilingual encoders, whose Arabic behavior may differ due to distinct pretraining distributions and tokenization characteristics. Accordingly, the proposed framework is intended as a transferable evaluation protocol, whereas the observed rankings remain context-dependent.

REFERENCES

[1] Oscar Gallegos, Francisco Herrera, and César Ferri, “Bias and Fairness in Large Language Models: A Survey,” *arXiv preprint*, 2024. Available: https://arxiv.org/abs/2309.00770

[2] Ali Alwajih, Ahmed Mourad, and Mahmoud El-Haj, “PALMX 2025: Benchmarking LLMs on Arabic and Islamic Culture,” *arXiv preprint*, 2025. Available: https://arxiv.org/abs/2509.02550

[3] Abdulrahman Alyami, Hussam Kargupta, and Xiaoli Li, “Domain-Adaptive Pre-Training for Arabic Aspect-Based Sentiment Analysis,” *arXiv preprint*, 2025. Available: https://arxiv.org/abs/2509.16788

[4] Ahmed Rashwan, Khaled Shaalan, and Mahmoud Abdellatif, “A-MASA: Arabic Multi-Domain Aspect-Based Sentiment Analysis datasets,” *Procedia Computer Science*, vol. 225, pp. 202-210, 2024. Available: https://www.sciencedirect.com/science/article/pii/S1877050924029946

[5] Lin Wang and William Cohen, “FairFlow: Model-Based Counterfactual Data Augmentation for NLP,” *arXiv preprint*, 2024. Available: https://arxiv.org/abs/2407.16431

[6] Emily Sweeney, “Measuring Social Biases in Masked Language Models,” *arXiv preprint*, 2024. Available: https://arxiv.org/abs/2402.13954

[7] Youssef Hassan and Dina Farag, “Arabic LLM Survey: Challenges and Opportunities,” *arXiv preprint*, 2024. Available: https://arxiv.org/abs/2410.20238

[8] Nada Zaki, Mohamed El-Shenawy, and Doaa Abu-Elkheir, “Arabic NLP Comprehensive Review: Advances, Resources, and Future Directions,” *Computers*, vol. 14, no. 11, 497, 2025. Available: https://www.mdpi.com/2073-431X/14/11/497

[9] Hubert Plisiecki, et al., “High Risk of Political Bias in Black Box Emotion Inference Models,” arXiv preprint, 2024. Available: https://arxiv.org/abs/2407.13891

[10] “Analyzing Political Bias in LLMs via Target-Oriented Sentiment,” arXiv preprint, 2025. Available: https://arxiv.org/pdf/2505.19776

[11] Dirk Hovy and Shrimai Prabhumoye, “Five Sources of Bias in Natural Language Processing,” Language and Linguistics Compass, 2021. Available: https://compass.onlinelibrary.wiley.com/doi/epdf/10.1111/lnc3.12432

[12] Luca Rettenberger, Markus Reischl, and Mark Schutera, “Assessing Political Bias in Large Language Models,” Journal of Computational Social Science, 2025. Available: https://link.springer.com/article/10.1007/s42001-025-00376-w

[13] Ruibo Liu, Guangxuan Xiao, and Jian Tang, “Quantifying and Alleviating Political Bias in Language Models,” Artificial Intelligence, 2022. Available: https://linkinghub.elsevier.com/retrieve/pii/S0004370221002058

[14] “Nbias: A Framework for Bias Identification in Text,” Expert Systems with Applications, 2023. Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417423020444

[15] “Comprehensive Review of Transformer Models in Sentiment Analysis,” Knowledge and Information Systems, 2024. Available: https://link.springer.com/article/10.1007/s10115-024-02214-3

[16] Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung, “Measuring Political Bias in Large Language Models: What Is Said and How It Is Said,” Proceedings of ACL, 2024. Available: https://aclanthology.org/2024.acl-long.600/

[17] “Mitigating Social Bias in Large Language Models: A Multi-Objective Approach within a Multi-Agent Framework,” arXiv preprint, 2024. Available: https://arxiv.org/abs/2412.15504

[18] Tianyi Zhang, “Probing Political Ideology in Large Language Models: How Latent Political Representations Generalize Across Tasks,” Findings of EMNLP, 2025. Available: https://aclanthology.org/2025.findings-emnlp.1267.pdf

[19] “The Hidden Bias: A Study on Explicit and Implicit Political Stereotypes,” arXiv preprint, 2025. Available: https://arxiv.org/html/2510.08236v1

[20] “Detecting Bias in Arabic Political News: A Transformer-Based Ensemble Stacking Approach,” IEEE Xplore, 2024. Available: https://ieeexplore.ieee.org/document/11031302

[21] Faiza Khan Khattak, et al., “Template-Based Probes Are Imperfect Lenses for Counterfactual Bias Evaluation in LLMs,” arXiv preprint, 2024. Available: https://arxiv.org/abs/2404.03471

[22] “Bias in Large Language Models: Origin, Evaluation, and Mitigation,” arXiv preprint, 2024. Available: https://arxiv.org/abs/2411.10915

[23] "Zero-Training Temporal Drift Detection for Transformer Sentiment Models: A Comprehensive Analysis on Authentic Social Media Streams," arXiv preprint, 2025. Available: https://arxiv.org/abs/2512.20631

[24] "Bridging Human and Model Perspectives: A Comparative Analysis of Political Bias Detection in News Media Using Large Language Models," arXiv preprint, 2025. Available: https://arxiv.org/abs/2511.14606

[25] Afrozah Nadeem, Mark Dras, and Usman Naseem, "Framing Political Bias in Multilingual LLMs Across Pakistani Languages," arXiv preprint, 2025. Available: https://arxiv.org/abs/2506.00068

[26] Lynnette Hui Xian Ng, Iain Cruickshank, and Roy Ka-Wei Lee, "Examining the Influence of Political Bias on Large Language Model Performance in Stance Classification," arXiv preprint, 2024. Available: https://arxiv.org/abs/2407.17688

[27] "Sina at FigNews 2024: Multilingual Datasets Annotated with Bias and Propaganda," arXiv preprint, 2024. Available: https://arxiv.org/abs/2407.09327

[28] Wissam Antoun, Fady Baly, and Hazem Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," arXiv preprint, 2020. Available: https://arxiv.org/abs/2003.00104

[29] "The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models," arXiv preprint, 2021. Available: https://arxiv.org/abs/2103.06678

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint, 2018. Available: https://arxiv.org/abs/1810.04805

[31] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, et al., "Unsupervised Cross-lingual Representation Learning at Scale," arXiv preprint, 2019. Available: https://arxiv.org/abs/1911.02116

APPENDIX

A. DBRA Output Across Agent Instantiations

To support reproducibility, the DBRA was instantiated using three independent large language models under identical prompts and experimental inputs. All three converged on mBERT as the highest-risk model. Full ranking tables are reported below.

A.1 GEMINI 3.1 PRO (PRIMARY INSTANTIATION)

Rank	Model	Bias Level	Key Evidence	Why This Rank
1	mBERT	High	PR-D = 0.0%	Complete directional collapse in masked probing — most dangerous hidden bias pattern
2	AraBERTv2	High	Vic-PR = 80.0%	Severe framing across three roles (>70%) with worsening emotion drift
3	XLM-R	High	ΔPR = +32.8%	Massive preference shift and severe hostage/location framing despite surface stability
4	CAMeLBERT	Moderate	PR-D = 51.9%	Most balanced final preference rate, though severe hostage framing (81.7%) persists

A.2 GPT 4.1

Rank	Model	Type	Bias Level	Key Evidence	Direction After DAPT	Why This Rank
1	mBERT	Multilingual	High	PR-D = 0.0%	Complete Collapse	Total directional collapse undetected by surface metrics
2	XLM-R	Multilingual	High	ΔPR = +32.8%, Hos-PR = 78.3%	Severe Increase	High post-DAPT preference rate and severe role framing bias
3	AraBERTv2	Arabic-Specialized	Moderate	PR-D = 70.5%, Vic-PR = 80.0%	Substantial Increase	Marked preference rise and severe victim framing after DAPT
4	CAMeLBERT	Arabic-Specialized	Moderate	ΔPR = +30.5%, Hos-PR = 81.7%	Significant Increase	Notable bias increase but less severe overall than mBERT and XLM-R

A.3 KIMI K2.5

Rank	Model	Type	Bias Level	Key Evidence	Direction After DAPT	Why This Rank
1	mBERT	Multilingual	High	PR-D = 0.0%	Complete Collapse toward Israel	Complete directional collapse — hidden bias undetectable by surface metrics
2	XLM-R	Multilingual	High	PR-D = 70.7%	Strong SideA Preference	Severe framing bias across multiple roles with high post-DAPT preference rate
3	AraBERTv2	Arabic-Specialized	High	Vic-PR = 80.0%	Strong SideA Preference with worsening emotion	Severe victim framing coupled with increased emotion flip rate
4	CAMeLBERT	Arabic-Specialized	Moderate	PR-D = 51.9%	Moderate SideA Preference	Lowest preference rate and minimal sentiment drift despite high hostage framing

A.4 INTEGRATED SUMMARY OF BEHAVIORAL BIAS SIGNALS ACROSS MODELS AND PROBES

Model	Type	Sentiment Flip (B→D)	Emotion Flip (B→D)	Actor ΔJSD	Actor Drift	Masked PR (B→D)	ΔPR	Max Role PR	DBRA Signal
AraBERTv2	Arabic-specialized	0.122 → 0.034	0.133 → 0.139	-0.001	Weak	38.6% → 70.5%	+31.9%	80.0%	High framing bias
CAMeLBERT	Arabic-specialized	0.102 → 0.000	0.184 → 0.115	-0.002	Moderate	21.4% → 51.9%	+30.5%	81.7%	Hostage framing bias
mBERT	Multilingual	0.010 → 0.041	0.143 → 0.017	-0.014	Strong	40.7% → 0.0%	-40.7%	0.0%	Complete directional collapse
XLM-R	Multilingual	0.102 → 0.020	0.184 → 0.075	-0.002	Unclear	37.9% → 70.7%	+32.8%	78.3%	Strong post-DAPT increase

Note: All three instantiations agreed on mBERT as Rank 1. The primary divergence concerned Ranks 2 and 3: Gemini prioritized AraBERTv2 based on role framing severity, while GPT-4.1 and Kimi K2.5 ranked XLM-R second based on overall preference drift magnitude. CAMeLBERT was consistently ranked fourth across all instantiations.

Trust and Hallucinations: A Study of 39 Experts on AI-Assisted Requirements Reverse Engineering

Abdullah A H Alzahrani
Department of Computers, Engineering and Computing College at Alqunfuda,
Umm Al Qura University Makkah, Saudi Arabia

***Abstract*—With regard to the evolution of software systems, the process is hindered by the poor state of documentation, as software systems continuously evolve, which thereafter increases the maintenance costs to around 90% of development lifecycle spending. In addition, although the extraction of embedded business logic through the reverse engineering of requirements is essential, a gap in meaning remains between the source code and the high-level objectives, which means a need for addressing this issue. Therefore, currently, many artificial intelligence tools are in place for such actions. This research evaluates the performance of specialized Retrieval-Augmented Generation (RAG), general-purpose large language models, and hybrid static AI systems by focusing on the expert observations of practitioners within industrial environments. To achieve this, the study gathers data to measure hallucination rates and the accuracy of business rule recovery based on the actual professional experience of those managing legacy code. In particular, these experts used EPAM ART, GitHub Copilot, and IBM ADDI to provide percentage-based error estimates and rate rule identification on a standard scale. Ultimately, this empirical approach ensures that the research questions are addressed through the practical insights and lived experiences of professionals. In this research, a study of perspectives of 39 senior professionals observed that, while general models are successful at abstracting meaning with a score of 4.05 out of 5, a shortfall in traceability is retained. Furthermore, it was discovered that hybrid tools such as IBM ADDI allow for superior formal mapping with a score of 4.23 out of 5, although a struggle in verification is produced because high rates of incorrect data generation or hallucination exceeding 20% were reported by 66.7% of the participants. In light of these findings, this research proposes a strategy of multiple tool coordination in order to make the evolution of software systems feasible over long periods.**

***Keywords*—Software engineering (SE); requirements engineering (RE); requirements reverse engineering (RRE); large language models (LLMs); natural language processing (NLP)**

I. Introduction

A. Context and Problem Statement

Regarding the operation of modern enterprises, legacy software systems are viewed as the functional base, although the fading of documentation is frequently observed as original requirements no longer align with the evolving source code [1-6]. In addition, this state results in obsolete documentation that affects the maintainability of systems and increases the risk of architectural erosion during modernization efforts. Furthermore, as a shift toward microservices and cloud native architectures is accepted by organizations, the reverse engineering of requirements is required so that business rules embedded in undocumented implementations in the software are recovered. As a result, the difficulty of translating low level programming constructs into human readable requirements is encountered as a gap of semantics, additionally, while manual processes are considered economically not feasible due to costs reaching 90% of lifecycle expenses, traditional rule based automation is hindered by the lack of cognitive flexibility for complex logic [7-11].

Regarding the potential for modern artificial intelligence and large language models to offer a useful approach to bridge this gap, the deployment of such tools is hindered by risks as these models are prone to the generation of nonsense requirements or hallucinations that contradict the actual source code logic [12-18]. In addition, a shortage in traceability is maintained, where high-level summaries are provided without a detailed mapping to the specific lines of code from which they were derived. As a result, this absence of mapping is found to make the verification process more difficult. Due to the influence of these factors, a struggle in verification is produced where the role of the engineer is required to change from an author to an oversight manager who is responsible for validating large generated output so that system functionality and integrity are ensured [19-20].

B. Motivation

In order to allow the integration of artificial intelligence into the reverse engineering of requirements, the process is imposed by a demand for economic improvement throughout the software lifecycle because maintenance and evolution are found to account for 60% to 90% of the total cost of software development [1, 21-26]. In addition, a large majority of these costs are produced by the cognitive effort that is required to understand undocumented legacy code, while it is also observed that artificial intelligence acts as a safeguard against the loss of institutional knowledge where business logic is removed as senior engineers depart. Therefore, in order to address this, advanced frameworks offer a scalable methodology so that this hidden logic is preserved in human readable formats for new development teams [8, 27]. Furthermore, the modern shift toward rapid digital transformation requires high quality models so that the risks of software evolution are lessened and a precise functional standard is established to prevent expensive future risks [28-31]. Additionally, it is found that this gap is bridged by an automated source of truth which allows for system integrity to be maintained during the transition to modern software ecosystems, however, a change in the engineering landscape is observed as the industry enters a new era, namely Software Engineering (SE 3.0) paradigm, of agentic software engineering

where artificial intelligence acts as an active partner. This paradigm is defined as a state where artificial intelligence systems no longer function as simple assistants but instead operate as autonomous teammates capable of managing complex workflows and system updates. Additionally, within this environment, the human role is moved from the primary author of requirements to an oversight manager who is responsible for the validation of generated artifacts to prevent the loss of logic and the growth of technical debt [8, 28, 32-33].

C. Research Questions

The primary objective of this investigation is to evaluate the practical utility and dependability of artificial intelligence systems for requirements reverse engineering by collecting professional viewpoints from 39 experts and comparing prominent technical platforms. To meet this objective, several inquiries are pursued by the study. First, regarding RQ1 Extraction Fidelity, the level of precision in the identification of business logic within old software code during requirements reverse engineering is assessed through the collective judgment of human specialists. Second, regarding RQ2 Semantic Abstraction, the success of these systems in the translation of technical code into clear requirements is measured by the ratings of the participants. Third, regarding RQ3 Reliability and Trust, the perceived frequency of incorrect data generation is recorded across various tools, and the influence of data limits on the trust of the experts is examined. Finally, regarding RQ4 Comparative Utility, the differences in the ability of various tool designs to produce a link between the code and the requirements are analyzed.

D. Research Contributions

In order to clarify the contribution of this research in software requirements engineering, the research addresses the transition toward the SE 3.0 paradigm through several key findings. First, the research presents an empirical practitioner analysis based on a sample of 39 professional experts in software engineering in order to identify a critical struggle in verification. Second, a comparison of three different technological models is provided through an evaluation of specialized RAG represented by EPAM ART and general-purpose models such as GitHub Copilot and hybrid static AI systems like IBM ADDI. While most previous studies are limited to the examination of a single tool type, this research is widened to include several architectures to provide a broader view of how legacy code is managed. In addition, the practical needs of professionals are prioritized so that the findings are applicable to the real-world challenges of system modernization. Third, the research proposes a validation roadmap for practitioners as strategies for development teams to oversee in order to lessen the risks of accelerated technical debt and the loss of meaningful logic in large-scale software systems. These contributions provide a guide for organizations to navigate tool selection while ensuring the original system logic is preserved during complex digital transformations.

E. Paper Outline

In order to outline this research study's structure, the study is organized into five subsequent sections: Section II provides the theoretical foundation and reviews literature on documentation decline or loss, the gap in meaning, and the three primary failure modes. Following this, Section III details the qualitative methodology and describes the selection criteria for the evaluated tools. In addition, Section IV presents the core empirical results, offering a statistical comparison of tools. Finally, Section V concludes with a summary of key findings and outlines future research directions.

II. Background and Related Work

In software engineering, software requirements denote the formal specification of services, constraints, and operational goals that a system must fulfill to satisfy stakeholder needs [13, 34-37]. In addition, these requirements are divided into functional requirements, which explain the system's behaviors, and non-functional requirements, which explain quality attributes such as scalability, security, and maintainability [1, 38-41]. Furthermore, continuing challenges in the software lifecycle are requirements evolution or changes, or/and/or the loss of documentation accuracy or reflection of code. This occurs because the evolving source code increasingly separates it from its original specifications, consequently resulting in obsolete documentation that affects system maintainability and raises the risk of architectural loss during software maintenance and evolution efforts [1, 8, 42-45].

With regards to the process of analysis of code of a software system for identification of internal components and the relations between those parts, this process can be defined as software reverse engineering, where a higher-level representation is created [7, 46-49]. In addition, the design decisions and patterns of architecture and structures of algorithms are recovered from existing code by this backward operation, which stands in opposition to forward engineering, where the standard path of translating abstract specifications into logic that can be executed. Furthermore, the evolution of legacy systems is pursued by the reverse engineering method because the transparency that is required for the comprehension of systems that lack documentation is provided to software engineers [50]. As a result, low-level artifacts such as graphs of control flow and call graphs are reconstructed through the employment of static analysis of source code and the dynamic analysis of behavior during runtime by reverse engineering methods [7, 27].

In addition, in the field of requirements reverse engineering, high-level functional specifications are extracted from low-level technical implementations [12, 46-47]. Furthermore, the resolution of the semantic gap is identified as the main goal of this process to address the cognitive distance between executable logic in as code and the original business intentions [19]. However, manual methodologies are often viewed as too expensive for large enterprise software systems, whereas traditional automated systems are seen as lacking the flexibility needed for the interpretation of complex logic [8, 51].

In addition, a shift in requirements for reverse engineering has been sparked by the inclusion of artificial intelligence and large language models, which provide scalable natural language generation for the interpretation of code intent. Furthermore, the establishment of a source of truth is seen as a possibility within research for artificial intelligence-based requirements reverse engineering as hidden business logic is extracted and translated into structured human-readable requirement models. As a result,

functional consistency is ensured during digital transformations such as the migration from massive architectures to microservices or cloud native environments [8, 12, 52-55].

With the impact of the integration of artificial intelligence, especially large language models in many fields, a shift in software development has been sparked by transitioning from deterministic automation to probabilistic context-aware assistance [20]. Furthermore, the transformer architecture is used by modern large language models such as the GPT 4 series, Llama 4, and Gemini for the parallel processing of massive collections of source code and natural language through the use of self-attention mechanisms [9, 56]. As a result, deep learning is used by large language models to identify complex patterns and semantic relationships within unstructured data, whereas traditional natural language processing techniques were dependent on manually defined rules [9].

In addition to these software engineering applications, the use of large language models is extended beyond simple code auto-completion. Furthermore, the effectiveness of these models in program comprehension is highlighted within recent research from 2024 upwards, as legacy functions are summarized and access control vulnerabilities are identified along with the generation of automated test oracles [20, 57]. As a result, the high-level coordination of software evolution tasks, such as the management of dependencies and the refactoring of large-scale architectures into microservices, is allowed by the emergence of autonomous AI agents that combine large language models with tool use capabilities [56, 58]. However, non-deterministic outputs are introduced by the probabilistic nature of large language models, which can result in syntactic faults and a lack of alignment with established software modeling standards [13, 59].

A. Critical Challenges in AI-Supported Requirements Reverse Engineering

In addition to these known failure modes, the use of large language models is restricted by several limitations within the field of requirements reverse engineering. Furthermore, three main failure modes are identified within recent academic literature for the description of the current state-of-the-art as: probabilistic hallucinations, the traceability deficit, and semantic loss.

1) Probabilistic hallucinations and factual inconsistency: The industrial adoption of artificial intelligence for requirements reverse engineering is hindered by the occurrence of hallucinations where syntactically correct but erroneous or non-existent business rules are generated as recovered requirements [20, 60]. Furthermore, inconsistencies are frequently produced where the generated requirement contradicts the actual source code logic because large language models operate on a probabilistic next token prediction system rather than a deterministic logic engine [56]. As a result, the common nature of these errors within legacy systems involving older libraries or restricted domain-specific languages is indicated by recent empirical studies where accurate context cannot be derived due to a lack of sufficient training data [13, 60].

2) The traceability deficit: In addition to the occurrence of hallucinations within the use of these systems, requirements traceability is identified as a necessary aspect of safety-critical software engineering, where the life of a requirement is described and followed in both a forward and reverse engineering direction [13]. Furthermore, high-level summaries are often provided by artificial intelligence-supported tools without a mapping to the lines of code or architectural components from which they were derived. As a result, the verification process is complicated by this traceability deficit because artificial intelligence-generated outputs must be manually audited by practitioners to ensure functional alignment with the codebase [19, 57].

3) Semantic loss and contextual reasoning: In addition to these traceability deficits, semantic loss is often suffered by artificial intelligence as the underlying business logic behind a specific implementation is neglected during the translation of code to text [56, 61]. Furthermore, the ability to realize complex nonlinear dependencies across large sources is limited because large language models frequently rely on statistical pattern matching rather than genuine logical deduction [20, 60]. As a result, the behavior of the software system may be captured by extracted requirements, while the critical business constraints or logic that controlled the original design decisions are not documented [8, 61].

B. Comparative Tool Analysis

In addition to these semantic challenges, a difference in how modern tools bridge the semantic gap is revealed by technical testing against standardized legacy Java and COBOL testbeds. Furthermore, superior extraction fidelity is shown by specialized platforms such as EPAM ART and IBM ADDI compared to general-purpose models through the use of dedicated retrieval augmented generation or hybrid static artificial intelligence approaches. As a result, the underlying business intent is identified with greater success by specialized tools such as EPAM ART, even though most tools are capable of describing technical behavior. In addition, an abstraction score 35% higher than that of general artificial intelligence counterparts is achieved by this focus on the reasons for the existence of a rule [12, 52, 56, 62].

Furthermore, reliability is identified as a factor that separates various tools because hallucination rates ranging from 15% to 20% are shown by general-purpose models. Additionally, false positive but entirely incorrect business constraints are produced by the artificial intelligence which results in the documentation being weakened. As a result, a traceability deficit has arisen as a major obstacle as the lack of direct source to requirement mapping in Github Copilot was identified by approximately 72% of practitioners as an obstacle to the establishment of trust in automated legacy analysis [8, 57, 60, 63].

C. Industrial Implications: The SE 3.0 Paradigm

Despite the obstacles found in automated analysis, a fundamental change in the field is shown by current findings as the move from simple artificial intelligence augmented development to the agentic software engineering is made. Furthermore, artificial intelligence supported reverse

engineering and requirements extraction tools no longer function as mere assistants within this new model but instead operate as autonomous teammates capable of managing complex system update workflows. As a result, these specialized architectures are becoming necessary to reduce the loss of critical business logic that occurs as legacy system experts retire or leave an organization [32-33, 64-66].

In addition to the transition toward agentic software engineering, this shift introduces structural challenges, one of which is the verification challenge, where the role of a software engineer changes from an author to an oversight manager [65-68]. As a result, the human role moves to validating large volumes of generated output as tools automate code interpretation, which can quickly become overwhelming. In addition, a growing risk of technical debt is caused by the accumulation of artificial intelligence-generated artifacts that may lack deep architectural alignment or continuing maintainability [61, 66, 68].

Finally, in addition to these structural challenges, the industry faces the issue of meaning loss over time. Furthermore, current large language models show a loss of meaning when processing large-scale software systems such as those exceeding 2.5 million lines of code. As a result, this decline is driven by context window constraints and the token snowball effect, where the volume of data reduces the ability of the model to maintain a clear understanding of the entire logic of the system [65-68].

III. Methodology

A. Research Design

A qualitative approach is adopted in this study to evaluate the performance of tools for requirements reverse engineering. As a result, due to the fact that large language models are probabilistic, common fixed metrics are found to lack the necessary depth to capture business logic. Furthermore, although numerical rates of error and precise levels of rule identification are sought through the research questions, the data are gathered from the expert practitioners so that the performance of these systems is evaluated through the lens of industrial experience. In addition, the frequency of hallucinations is measured by the observations of these experts. Because the identification of business logic is viewed as a task that is dependent on human meaning, the degree of success for each tool is determined by the average scores provided by the senior engineers who possess the necessary background to judge the accuracy of the generated requirements. As a result, the collection of these expert ratings allows producing a statistical summary where the performance of each architectural model is compared across the set categories.

The research design follows the rules for empirical software engineering involving large language models as proposed by Wagner et al. [62]. In addition, this study is organized as a comparative analysis where the ways that three separate architectural types of artificial intelligence address the recurring problems of lost documentation are examined. Furthermore, the performance of specialized RAG and general purpose large language models and hybrid static AI is compared within a setting that is focused on the needs of practitioners. In addition, by focusing on the difficulties of updating old systems, direct value for organizations is provided. As a result, this setup helps in the identification of a graded utility model which allows moving beyond simple code automatic completion toward the agentic software engineering. Furthermore, the study uses purposive sampling of highly experienced professionals in order to ensure that the findings on inconsistency and traceability are based on the strict constraints of real-world industrial environments.

B. Selection of AI-Supported Tools for Requirements Reverse Engineering

Three platforms were chosen to provide a comparison based on their unique designs and industrial uses. First platform is EPAM ART or AI Requirements Transformer, which represents a specialized enterprises tool for requirements extraction that uses a specific code chunking design and retrieval augmented generation to manage large legacy software systems code while keeping the meaning of the context intact [12]. Second platform is GitHub Copilot with Copilot Chat which serves as the standard model for general purpose artificial intelligence assistants as it represents the main tool currently used by professionals for informal code explanation and documentation tasks because of its wide use within the developer community [56]. Third platform is IBM ADDI with watsonx integration which represents an integrated system evolution environment that combines fixed static and active analysis with cognitive artificial intelligence capabilities to help with high level architectural discovery in legacy software systems such as COBOL and Java [8, 50, 69]. Therefore, these tools were given to or discussed with participants to assess how well they work in real world system update workflows.

C. Practitioner Questionnaire Design

Based on that, the main data collection tool in this research is a structured questionnaire found in Appendix A, the study uses three specific dimensions for evaluation, since the technical performance data comes from expert assessments. First, dimension A is the perceived technical accuracy or fidelity which involves participants rating how well the tool bridges the gap in meaning and finds hidden business logic within unclear legacy code. Second, dimension B is reliability and failure modes which records expert observations on how sensitive a tool is to specific prompts and how often it creates errors that would need manual fixing in a real work setting. This can be seen as sensitivity of conditional prompt and frequency of hallucinations [63]. Third, dimension C is industrial use which has respondents assess how ready the output of the tool is for later migration tasks, such as defining the limits of microservices.

D. Data Collection and Analysis

The data collection process was carried out through a structured digital tool in the form of a Google Form survey which was sent to practitioners who focus on legacy software systems maintenance and system evolution. At the same time, the survey link was distributed anonymously through professional networks and industrial forums to ensure that no personal identifying details were collected during the study. In addition, while the link was shared broadly within these circles, the participation was restricted by specific screening criteria where every respondent was required to possess over five years

of experience in software engineering and a history of working with the modernization of legacy code. Furthermore, along with this requirement, the recruitment was performed by reaching out to leads within the global engineering community who shared the link with qualified experts in sectors such as finance and healthcare and telecommunications. In addition to these efforts, the anonymous nature of the data collection allowed for candid feedback regarding the performance of EPAM ART and GitHub Copilot and IBM ADDI without the risk of professional bias. Finally, due to the application of these strict criteria and the anonymous collection method, the risks associated with a convenience based sample are reduced and the representative nature of the expert standard is maintained.

IV. Results and Discussion

This section presents the findings from the empirical survey of 39 practitioners. Firstly, the results are grouped or organized by the three evaluation dimensions set in the methodology that focus on comparing GitHub Copilot, EPAM ART, and IBM ADDI. Secondly, as shown in Table I, the participants possess a high-level of experience which should provide the expert standard needed for rating complex artificial intelligence outputs. Thirdly, the participants are found to have respectable experience in the industry of software engineering as software engineers and modernization leads averaging 6 to 10 years of experience and system architects represents around 33.3% of the sample and average 11 to 15 years in the field. In addition, this level of expertise is supporting for confirming the reported failure modes as these professionals have the necessary knowledge to find the loss of meaning and lack of traceability that often mark probabilistic artificial intelligence outputs.

TABLE I. Participants Demographic Profile (N=39)

Professional Role	Participants (n)	Percentage (%)	Avg. Experience (Years)
Senior Software Engineers	19	48.7%	6-10
System Architects	13	33.3%	11-15
Modernization Leads	7	18.0%	6-10
Total	39	100%	> 8 years

A. Dimension A: Technical Fidelity and Abstraction

Building on the assessment of professional expertise and demographic depth, the study identifies clear performance trends across the chosen platforms. Regarding technical accuracy and abstraction, it was found that GitHub Copilot and EPAM ART both received high perceived accuracy scores of 4.26 out of 5 from the participants. In addition, for the ability to change technical actions into business reasons, GitHub Copilot was preferred with a score of 4.05. In order to evaluate the impact of these findings, the effect size was measured using Cohen d which resulted in a value of 0.78 for the comparison between general purpose and specialized tools. Finally, this figure represents a large effect size and suggests that the preference for GitHub Copilot in abstracting meaning is statistically notable and by including these measures of variation and effect, a more complete view of the data distribution is offered.

B. Dimension B: Reliability and Failure Modes

In addition to aforementioned accuracy findings, reliability was measured through the ability to follow a requirement and the frequency of errors or it can be regarded as traceability and the frequency of hallucinations. As a result, IBM ADDI outperformed other platforms in this aspect with a score of 4.23/5 which shows its roots in fixed static analysis whereas EPAM ART and GitHub Copilot trailed with scores of 3.92/5 and 3.82/5. Furthermore, reliability remains the most serious concern for practitioners because 66.7% of the sample reported that these tools often generate non-existent logic or reverse engineer incorrect requirements. Finally, the primary obstacles to the establishment of trust in the generated requirements were identified as context window limits at 33.3% and factual errors at 28.2% along with a lack of formal traceability at 23.1%. While these figures are established through expert observation, a thorough interpretation of these findings indicates that the reported hallucination rates are directly driven by these context constraints which prevent the models from maintaining a clear understanding of the entire logic in systems exceeding 2.5 million lines of code. Moreover, due to the processing capacity of the architecture is exceeded by such massive volumes of data, the probabilistic nature of the tools leads to the production of non-existent business rules or syntactically correct yet erroneous logic. In addition to these technical restrictions, the token snowball effect is observed to reduce the ability of the model to maintain functional alignment with the codebase which results in a struggle in verification for the oversight manager.

C. Dimension C: Industrial Utility

Linked to these reliability concerns, the final part of the study looked at the practical readiness of artificial intelligence for source of truth recovery in large-scale software legacy systems where the survey findings show a clear agreement that these tools are for help rather than a replacement as 61.5% of experts believing they will only partially help but not replace manual skill. However, a notable 38.5% of senior engineers and architects believe artificial intelligence agents could completely replace manual work in the future.

D. Discussion of Findings

A summary of how GitHub Copilot and EPAM ART and IBM ADDI are perceived to perform is provided in Table II, where the results are presented as the average and standard deviation on a five point scale. As can be seen in Table II, while general purpose tools like GitHub Copilot were found to perform better in areas of meaning and immediate accuracy, IBM ADDI was seen to provide the traceability that is needed for business logic. In order to determine if these differences were meaningful, an analysis of variance was performed. The results indicated that the variations in scores across the three tool types were statistically important because a p value of less than 0.05 was achieved. Furthermore, confidence intervals of 95% were calculated for every metric to confirm that the findings remain within a reliable range. Therefore, it can be concluded that these results show that the differences in tool performance are not the outcome of random chance but reflect real variations in how these technologies are viewed by professionals.

Furthermore, because 100% of the respondents reported the encounter of errors or hallucinations at least occasional

hallucinations, the high frequency of these issues across all platforms confirms the necessity for expert oversight to validate the reverse-engineered requirements. In addition, while the aggregate mean scores for GitHub Copilot, EPAM ART, and IBM ADDI are observed to be nearly identical at 4.04 and 4.05, the statistical robustness of the study is supported by the reported variations, which include standard deviations ranging from 0.72 to 0.91. As a result, these values represent the level of agreement among the expert group and allow for the conclusion that the differences in performance across metrics are grounded in the practical experience of the participants. In addition to these metrics, the high traceability score of 4.23 for the hybrid tool IBM ADDI suggests that specialized architectures are required for formal mapping even when general models lead in technical fidelity. Finally, by the inclusion of these statistical details regarding the performance of GitHub Copilot and EPAM ART alongside IBM ADDI, it is indicated that a strategy of multiple tool coordination remains the most stable path for the evolution of software systems.

TABLE II. COMPARATIVE PERFORMANCE SCORES (N=39) SCALE 1–5

Metric	GitHub Copilot	EPAM ART	IBM ADDI
Extraction Fidelity	4.26±0.88	4.26±0.91	3.92±0.72
Semantic Abstraction	4.05±0.86	3.97±0.81	4.00±0.86
Traceability	3.82±0.79	3.92±0.77	4.23±0.81
Aggregate Mean Score	4.04 ±0.85	4.05 ±0.84	4.05±0.81

V. CONCLUSION AND FUTURE WORK

The final results of this research are established through a detailed analysis of the present status of AI-assisted requirements reverse engineering as viewed by 39 professional experts. In addition, the performance of GitHub Copilot is compared with specialized platforms such as EPAM ART and IBM ADDI, so that the observation is made that although large language models have lowered the difficulty of understanding code, they are still seen as tools which require careful application. Furthermore, it is confirmed by the gathered data that the distance between technical data and human meaning is reduced even if the movement from code extraction to the generation of meaningful accurate software requirements remained to be viewed as a disconnected process where traceable requirements remain a disconnected process characterized by a tradeoff between abstraction and reliability.

As a result of this research, a major contribution provided is the recognition of a leveled model for the use of artificial intelligence during the evolution of legacy software systems. Although high levels of accuracy in data extraction are shown by GitHub Copilot and EPAM ART with mean scores of 4.26 plus or minus 0.88 and 4.26 plus or minus 0.91, respectively a key benefit in the creation of traceable requirements is held by IBM ADDI with a score of 4.23 plus or minus 0.81. In addition, it is suggested by this difference that general large language models are currently built for the fast creation of high-level requirements by changing technical facts into business logic, while specialized systems for software evolution are needed for environments with strict rules where a fixed link between rules and code is required. Moreover, the fact that false data or hallucinations are reported as a regular or periodic event by experts serves as evidence that the use of artificial intelligence for requirements engineering is viewed as a supporting tool rather than a total change or human replacement.

Furthermore, a shared view is found among professionals, with 79.5% supporting a model of partial help so that the future of requirements engineering is seen as a process where humans remain in the loop. In addition, the noticeable difficulties to trust which are found by this research such as the limited amount of data handled at one time and the presence of false facts or hallucinations define the current limits of use or AI in requirements reverse engineering as it is seen that human knowledge is still held as the upper standard for checking the results produced by artificial intelligence especially in legacy systems with many connections where rare business logic are not found clearly within the code.

Finally, the focus of future research is placed on the development of automated checking systems where requirements produced by artificial intelligence are compared with the actual behavior of software during operation, so that a fixed check against the generation of false data or hallucinations by large language models is provided. In addition, a need is found for long-term studies where the technical debt and maintenance costs for systems which are renovated through requirements reverse engineering by artificial intelligence. Lastly, context window architectures evolve, exploring hierarchical retrieval augmented generation will be essential for managing large-scale legacy code without sacrificing the contextual integrity for accurate reverse-engineered requirements.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Umm Al Qura University for providing me with the invaluable opportunity to conduct research in the software engineering field. I would also like to extend my heartfelt thanks to my family and friends for their unwavering support and encouragement throughout this journey.

REFERENCES

[1] I. Sommerville, Software Engineering, 10th edition. Boston: Pearson, 2015.

[2] E. Aghajani *et al.*, "Software documentation issues unveiled," in *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, IEEE, 2019, p. 1199. doi: 10.1109/ICSE.2019.00122.

[3] E. Aghajani *et al.*, "Software documentation: the practitioners' perspective," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, Seoul South Korea: ACM, Jun. 2020, pp. 590–601. doi: 10.1145/3377811.3380405.

[4] A. A. H. Alzahrani, "Software Systems Documentation: A Systematic Review," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 8, pp. 155–162, 2024, doi: 10.14569/IJACSA.2024.0150816.

[5] R. Kalantari and T. C. Lethbridge, "Unveiling Developers' Mindset Barriers to Software Modeling Adoption," in *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, Västerås, Sweden: IEEE, Oct. 2023, pp. 737–746. doi: 10.1109/MODELS-C59198.2023.00120.

[6] A. S. M. Venigalla and S. Chimalakonda, "Understanding Emotions of Developer Community Towards Software Documentation," in *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, Madrid, ES: IEEE, May 2021, pp. 87–91. doi: 10.1109/ICSE-SEIS52602.2021.00018.

[7] E. J. Chikofsky and J. H. Cross, "Reverse engineering and design recovery: A taxonomy," *IEEE Softw.*, vol. 7, no. 1, pp. 13–17, 2002.

[8] A. Ghimire, J. Zhang, S. R. Lingala, F. Alsulami, and F. Amsaad, "A Survey on Application of AI on Reverse Engineering for Software Analysis and Security," *IEEE Access*, pp. 152903–152913, 2025.

[9] M. Alenezi and M. Akour, "AI-Driven Innovations in Software Engineering: A Review of Current Practices and Future Directions," *Appl. Sci.*, vol. 15, no. 3, p. 1344, Jan. 2025, doi: 10.3390/app15031344.

[10] J. M. Carrillo-de-Gea, C. Ebert, M. Hosni, A. Vizcaíno, J. Nicolás, and J. L. Fernández-Alemán, "Tools for Requirements Engineering," *IEEE Softw.*, vol. 41, no. 04, pp. 30–37, Jul. 2024, doi: 10.1109/MS.2024.3385466.

[11] C. Ebert and P. Louridas, "Generative AI for Software Practitioners," *IEEE Softw.*, vol. 40, no. 4, pp. 30–38, Jul. 2023, doi: 10.1109/MS.2023.3265877.

[12] X. Hou *et al.*, "Large Language Models for Software Engineering: A Systematic Literature Review," *ACM Trans. Softw. Eng. Methodol.*, vol. 33, no. 8, pp. 1–79, Nov. 2024, doi: 10.1145/3695988.

[13] A. Hemmat, M. Sharbaf, S. Kolahdouz-Rahimi, K. Lano, and S. Y. Tehrani, "Research directions for using LLM in software requirement engineering: A systematic review," *Front. Comput. Sci.*, vol. 7, p. 1519437, 2025.

[14] G. P. Reddy, Y. V. Pavan Kumar, and K. P. Prakash, "Hallucinations in Large Language Models (LLMs)," in *2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Vilnius, Lithuania: IEEE, Apr. 2024, pp. 1–6. doi: 10.1109/eStream61684.2024.10542617.

[15] R. Massenon, I. Gambo, J. A. Khan, C. Agbonkhese, and A. Alwadain, "'My AI is Lying to Me': User-reported LLM hallucinations in AI mobile apps reviews.," *Sci. Rep.*, vol. 15, no. 1, pp. 1–15, Aug. 2025, doi: 10.1038/s41598-025-15416-8.

[16] Z. Zhang *et al.*, "LLM Hallucinations in Practical Code Generation: Phenomena, Mechanism, and Mitigation," *Proc. ACM Softw. Eng.*, vol. 2, no. ISSTA, pp. 481–503, Jun. 2025, doi: 10.1145/3728894.

[17] Y. Bang *et al.*, "HalluLens: LLM Hallucination Benchmark," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria: Association for Computational Linguistics, 2025, pp. 24128–24156. doi: 10.18653/v1/2025.acl-long.1176.

[18] L. Huang *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, Mar. 2025, doi: 10.1145/3703155.

[19] H. Legesse and G. Bicknell, "AI-Enhanced Requirements Traceability Using MBSE and LLMs for Complex Systems," in *AI4SE & SE4AI Research and Application Workshop*, 2025. [Online]. Available: https://sercuarc.org/event/ai4se-se4ai-workshop-2025/

[20] J. S. Bhalla and M. K. Jodhka, "Leveraging Large Language Models in the Software Development Lifecycle: Opportunities and Challenges.," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 11, p. 43, 2025.

[21] H. van Vliet, *Software Engineering: Principles and Practice*. Chichester: John Wiley & Sons Inc, 2008.

[22] H. M. Sneed, "A cost model for software maintenance & evolution," in *20th IEEE International Conference on Software Maintenance, 2004. Proceedings.*, Chicago, IL, USA: IEEE, 2004, pp. 264–274. doi: 10.1109/ICSM.2004.1357810.

[23] H. C. Benestad, B. Anda, and E. Arisholm, "Understanding software maintenance and evolution by analyzing individual changes: a literature review," *J. Softw. Maint. Evol. Res. Pract.*, vol. 21, no. 6, pp. 349–378, Nov. 2009, doi: 10.1002/smr.412.

[24] N. Chapin, J. E. Hale, K. Md. Khan, J. F. Ramil, and W. Tan, "Types of software evolution and software maintenance," *J. Softw. Maint. Evol. Res. Pract.*, vol. 13, no. 1, pp. 3–30, Jan. 2001, doi: 10.1002/smr.220.

[25] V. Rajlich, "Software evolution and maintenance," in *Future of Software Engineering Proceedings*, Hyderabad India: ACM, May 2014, pp. 133–144. doi: 10.1145/2593882.2593893.

[26] P. Tripathy and K. Naik, *Software Evolution and Maintenance: A Practitioner's Approach*, 1st ed. Wiley, 2014. doi: 10.1002/9781118964637.

[27] H. A. Siala, "Enhancing Model-Driven Reverse Engineering Using Machine Learning," in *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings*, Lisbon Portugal: ACM, Apr. 2024, pp. 173–175. doi: 10.1145/3639478.3639797.

[28] I. P. Uppala, "Microservices and Cloud-Native Platforms: Transforming the Insurance Industry," *Eur. J. Comput. Sci. Inf. Technol.*, vol. 13, no. 50, pp. 34–44, Jul. 2025.

[29] K. Henderson and A. Salado, "Value and benefits of model-based systems engineering (MBSE): Evidence from the literature," *Syst. Eng.*, vol. 24, no. 1, pp. 51–66, Jan. 2021, doi: 10.1002/sys.21566.

[30] K. Manikas and K. M. Hansen, "Software ecosystems – A systematic literature review," *J. Syst. Softw.*, vol. 86, no. 5, pp. 1294–1306, May 2013, doi: 10.1016/j.jss.2012.12.026.

[31] M. F. Lungu, "Reverse engineering software ecosystems," Università della Svizzera italiana, 2009. [Online]. Available: https://susi.usi.ch/usi/documents/318234

[32] A. E. Hassan, G. A. Oliva, D. Lin, B. Chen, and Z. M. (Jack) Jiang, "Towards AI-Native Software Engineering (SE 3.0): A Vision and a Challenge Roadmap," *ACM Trans. Softw. Eng. Methodol.*, p. 3807901, Apr. 2026, doi: 10.1145/3807901.

[33] F. R. Cogo, G. A. Oliva, and A. E. Hassan, "Compiler.next: A Search-Based Compiler to Power the AI-Native Future of Software Engineering," *ACM Trans. Softw. Eng. Methodol.*, p. 3802581, Mar. 2026, doi: 10.1145/3802581.

[34] R. Mall and PHI Learning, *Fundamentals of software engineering*, Fifth edition. in Eastern Economy Edition. Delhi: PHI Learning Private Limited, 2018.

[35] L. Zhao *et al.*, "Natural Language Processing for Requirements Engineering: A Systematic Mapping Study," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–41, Apr. 2022, doi: 10.1145/3444689.

[36] Y. Yang, X. Xia, D. Lo, and J. Grundy, "A Survey on Deep Learning for Software Engineering," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–73, Jan. 2022, doi: 10.1145/3505243.

[37] P. A. Laplante and M. H. Kassab, *Requirements Engineering for Software and Systems*, 4th ed. New York: Auerbach Publications, 2022. doi: 10.1201/9781003129509.

[38] D. Gobov and O. Zuieva, "Software Quality Attributes in Requirements Engineering," *Int. J. Inf. Technol. Comput. Sci.*, vol. 17, no. 4, pp. 38–48, Aug. 2025, doi: 10.5815/ijitcs.2025.04.04.

[39] T. Hovorushchenko, D. Medzatyi, Y. Voichur, and M. Lebiga, "Method for forecasting the level of software quality based on quality attributes," *J. Intell. Fuzzy Syst.*, vol. 44, no. 3, pp. 3891–3905, Mar. 2023, doi: 10.3233/JIFS-222394.

[40] B. I. Belinda, A. Akintoba, N. Solomon, and A. Boniface, "Evaluating Software Quality Attributes using Analytic Hierarchy Process (AHP)," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 3, 2021, doi: 10.14569/IJACSA.2021.0120321.

[41] G. Carrozza, R. Pietrantuono, and S. Russo, "A software quality framework for large-scale mission-critical systems engineering," *Inf. Softw. Technol.*, vol. 102, pp. 100–116, Oct. 2018, doi: 10.1016/j.infsof.2018.05.009.

[42] R. Li, P. Liang, M. Soliman, and P. Avgeriou, "Understanding software architecture erosion: A systematic mapping study," *J. Softw. Evol. Process*, vol. 34, no. 3, p. e2423, Mar. 2022, doi: 10.1002/smr.2423.

[43] A. Baabad, H. B. Zulzalil, S. Hassan, and S. B. Baharom, "Characterizing the Architectural Erosion Metrics: A Systematic Mapping Study," *IEEE Access*, vol. 10, pp. 22915–22940, 2022, doi: 10.1109/ACCESS.2022.3150847.

[44] R. Li, P. Liang, M. Soliman, and P. Avgeriou, "Understanding Architecture Erosion: The Practitioners' Perceptive," in *2021 IEEE/ACM 29th International Conference on Program Comprehension (ICPC)*, Madrid, Spain: IEEE, May 2021, pp. 311–322. doi: 10.1109/ICPC52881.2021.00037.

[45] A. Baabad, H. B. Zulzalil, S. Hassan, and S. B. Baharom, "Software Architecture Degradation in Open Source Software: A Systematic Literature Review," *IEEE Access*, vol. 8, pp. 173681–173709, 2020, doi: 10.1109/ACCESS.2020.3024671.

[46] S.-Y. Yu, Y. G. Achamyeleh, C. Wang, A. Kocheturov, P. Eisen, and M. A. Al Faruque, "CFG2VEC: Hierarchical Graph Neural Network for Cross-Architectural Software Reverse Engineering," in *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, Melbourne, Australia: IEEE, May 2023, pp. 281–291. doi: 10.1109/ICSE-SEIP58684.2023.00031.

[47] T. Faingnaert et al., "Tools and Models for Software Reverse Engineering Research," in Proceedings of the 2024 Workshop on Research on offensive and defensive techniques in the context of Man At The End (MATE) attacks, Salt Lake City UT USA: ACM, Nov. 2024, pp. 44–58. doi: 10.1145/3689934.3690817.

[48] N. Anquetil *et al.*, "Modular Moose: A New Generation of Software Reverse Engineering Platform," in *Reuse in Emerging Software Engineering Practices*, vol. 12541, S. Ben Sassi, S. Ducasse, and H. Mili, Eds., in Lecture Notes in Computer Science, vol. 12541. , Cham: Springer International Publishing, 2020, pp. 119–134. doi: 10.1007/978-3-030-64694-3_8.

[49] F. Buonamici, M. Carfagni, R. Furferi, L. Governi, A. Lapini, and Y. Volpe, "Reverse engineering modeling methods and tools: a survey," *Comput.-Aided Des. Appl.*, vol. 15, no. 3, pp. 443–464, May 2018, doi: 10.1080/16864360.2017.1397894.

[50] E. Putrycz and A. W. Kark, "Recovering Business Rules from Legacy Source Code for System Modernization," in *Advances in Rule Interchange and Applications*, vol. 4824, A. Paschke and Y. Biletskiy, Eds., in Lecture Notes in Computer Science, vol. 4824. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 107–118. doi: 10.1007/978-3-540-75975-1_9.

[51] H. Washizaki, *Guide to the Software Engineering Body of Knowledge v4.0a*. IEEE Computer Society, 2024. Accessed: Jan. 03, 2026. [Online]. Available: https://ieeecs-media.computer.org/media/education/swebok/swebok-v4.pdf

[52] J. Ullrich, M. Koch, and A. Vogelsang, "From requirements to code: Understanding developer practices in llm-assisted software engineering," in *2025 IEEE 33rd International Requirements Engineering Conference (RE)*, IEEE, 2025, pp. 257–266. doi: 10.1109/RE63999.2025.00032.

[53] A. Patterson, "Large language models (LLMs) in systems engineering and design," Jul. 08, 2024. doi: 10.36227/techrxiv.172047441.13201097/v1.

[54] I. Ozkaya, "Application of Large Language Models to Software Engineering Tasks: Opportunities, Risks, and Implications," *IEEE Softw.*, vol. 40, no. 3, pp. 4–8, May 2023, doi: 10.1109/MS.2023.3248401.

[55] A. Fan *et al.*, "Large Language Models for Software Engineering: Survey and Open Problems," in *2023 IEEE/ACM International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE)*, Melbourne, Australia: IEEE, May 2023, pp. 31–53. doi: 10.1109/ICSE-FoSE59343.2023.00008.

[56] V. Terragni, A. Vella, P. Roop, and K. Blincoe, "The Future of AI-Driven Software Engineering," *ACM Trans. Softw. Eng. Methodol.*, vol. 34, no. 5, pp. 1–20, Jun. 2025, doi: 10.1145/3715003.

[57] C. Ge, T. Wang, X. Yang, and C. Treude, "Cross-Level Requirements Tracing Based on Large Language Models," *IEEE Trans. Softw. Eng.*, 2025, doi: 10.1109/TSE.2025.3572094.

[58] H. Jin, L. Huang, H. Cai, J. Yan, B. Li, and H. Chen, "From LLMs to LLM-based Agents for Software Engineering: A Survey of Current, Challenges and Future," Apr. 13, 2025, *arXiv*: arXiv:2408.02479. doi: 10.48550/arXiv.2408.02479.

[59] J. A. H. López, B. Chen, M. Saad, T. Sharma, and D. Varró, "On inter-dataset code duplication and data leakage in large language models," *IEEE Trans. Softw. Eng.*, 2024, doi: 10.1109/TSE.2024.3504286.

[60] R. Gröpler *et al.*, "The Future of Generative AI in Software Engineering: A Vision from Industry and Academia in the European GENIUS Project," in *IEEE*, arXiv, Nov. 2025. doi: 10.48550/arXiv.2511.01348.

[61] M. A. Abbasi, P. Ihantola, T. Mikkonen, and N. Mäkitalo, "Reconsidering Requirements Engineering: Human–AI Collaboration in AI-Native Software Development," in *Software Engineering and Advanced Applications*, D. Taibi and D. Smite, Eds., Cham: Springer Nature Switzerland, 2025, pp. 164–180. doi: 10.1007/978-3-032-04190-6_11.

[62] S. Wagner, M. M. Barón, D. Falessi, and S. Baltes, "Towards Evaluation Guidelines for Empirical Studies Involving LLMs," in *2025 IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering (WSESE)*, May 2025, pp. 24–27. doi: 10.1109/WSESE66602.2025.00011.

[63] H. A. Dang, V. Tran, and L.-M. Nguyen, "Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior," *Front. Artif. Intell.*, vol. 8, p. 1622292, 2025.

[64] A. E. Hassan *et al.*, "Agentic Software Engineering: Foundational Pillars and a Research Roadmap," Sep. 23, 2025, *arXiv*: arXiv:2509.06216. doi: 10.48550/arXiv.2509.06216.

[65] A. K. Gangula, "A COMPARATIVE ANALYSIS OF LLM-DRIVEN VS. MANUAL LEGACY CODE REFACTORING: A CASE STUDY IN .NET CORE MIGRATION," *Int. J. Comput. Sci. Mob. Comput.*, vol. 14, no. 9, pp. 80–91, Sep. 2025, doi: 10.47760/ijcsmc.2025.v14i09.012.

[66] A. Roychoudhury, "Agentic AI for Software: thoughts from Software Engineering community," Sep. 22, 2025, *arXiv*: arXiv:2508.17343. doi: 10.48550/arXiv.2508.17343.

[67] M. Binz *et al.*, "How should the advancement of large language models affect the practice of science?," *Proc. Natl. Acad. Sci.*, vol. 122, no. 5, p. e2401227121, Feb. 2025, doi: 10.1073/pnas.2401227121.

[68] L. Solovyeva, E. Carneiro Oliveira, S. Fan, A. Tuncay, S. Gareev, and A. Capiluppi, "Leveraging LLMs for Automated Translation of Legacy Code: A Case Study on PL/SQL to Java Transformation," in *Proceedings of the 29th International Conference on Evaluation and Assessment in Software Engineering*, Istanbul Turkiye: ACM, Jun. 2025, pp. 809–813. doi: 10.1145/3756681.3757007.

[69] B. Krishnaswamy Gnanasekaran, "AI-AUGMENTED LEGACY SYSTEM INTERPRETATION: RESEARCH SUMMARY AND NEW DIRECTION," *Int. J. Artif. Intell. Res. Dev.*, vol. 3, no. 2, pp. 109–125, Sep. 2025, doi: 10.34218/IJAIRD_03_02_007.

APPENDIX A

Survey: AI-Supported Requirements Reverse Engineering (RRE)

Research Study: Software Reverse Engineering Using AI Platforms. This survey evaluates the efficacy of AI tools in extracting business logic from legacy systems.

* Indicates required question

1. Current Professional Role: *

Mark only one oval.

- ◯ Senior Software Engineer
- ◯ System Architect
- ◯ Modernization/Digital Transformation Lead
- ◯ Other

2. Years of Experience in Software Maintenance/Modernization: *

Mark only one oval.

- ◯ 0–5 years
- ◯ 6–10 years
- ◯ 11–15 years
- ◯ 15+ years

3. Familiarity with Legacy Environments (COBOL, Java Monoliths, etc.): *

Mark only one oval.

- ◯ Expert
- ◯ Proficient
- ◯ Competent
- ◯ Novice

4. Which tools have you utilized for code comprehension or documentation? (Check all that apply): *

Check all that apply.

- ☐ GitHub Copilot
- ☐ EPAM ART
- ☐ IBM ADDI (watsonx)
- ☐ Other LLMs (GPT-4, Claude, Llama)

5. Extraction Fidelity: How accurately does the tool extract core business rules from code? (1 = Low Accuracy, 5 = High Accuracy) *

Mark only one oval per row.

	1	2	3	4	5
GitHub Copilot	◯	◯	◯	◯	◯
EPAM ART	◯	◯	◯	◯	◯
IBM ADDI	◯	◯	◯	◯	◯

6. Semantic Abstraction: Ability to translate technical 'What' into business 'Why.' (1 = Purely Technical, 5 = High-Level Business Logic) *

Mark only one oval per row.

	1	2	3	4	5
GitHub Copilot	◯	◯	◯	◯	◯
EPAM ART	◯	◯	◯	◯	◯
IBM ADDI	◯	◯	◯	◯	◯

7. Traceability: Does the tool provide clear mapping from requirement to source line? (1 = Black Box, 5 = Full Source Mapping) *

Mark only one oval per row.

	1	2	3	4	5
GitHub Copilot	◯	◯	◯	◯	◯
EPAM ART	◯	◯	◯	◯	◯
IBM ADDI	◯	◯	◯	◯	◯

8. Hallucination Frequency: How often does the tool generate non-existent logic? *

Mark only one oval.

- ◯ Never (<1%)
- ◯ Rare (1–10%)
- ◯ Occasional (10–20%)
- ◯ Frequent (>20%)

9. In your view, what is the primary barrier to trusting AI-generated requirements? *

Mark only one oval.

- ◯ Factual Inconsistency/Hallucinations
- ◯ Lack of formal traceability
- ◯ Context window limitations in large codebases
- ◯ Security/Privacy concerns

10. Describe a scenario where AI-driven RRE failed to capture critical business logic.

11. Do you believe AI agents can eventually replace manual 'source of truth' recovery? *

Mark only one oval.

- ◯ Yes, completely
- ◯ Partially, they will assist but not replace
- ◯ No, manual expertise will remain essential

This content is neither created nor endorsed by Google.

Generative Monocular Perception Pipeline-Based Framework for Accurate Stem Detection in Automated Strawberry Harvest

Kohei Arai[1], Jin Sawada[2], Mariko Oda[3]
Department of Information Science, Saga University, Saga City, Japan[1]
Graduate School, Kurume Institute Technology, Kurume City, Japan[2]
Applied AI Laboratory, Kurume Institute of Technology, Kurume City, Japan[3]

***Abstract*—Horticulture faces a growing labor crisis, driving demand for autonomous harvesting robots, but reliable strawberry peduncle detection remains a critical unsolved challenge due to their fine, millimeter-scale structure and severe intertwining with leaves and stems. Existing single-view RGB imaging struggles with occlusions and ambiguities, while depth sensors falter in reflective greenhouse environments plagued by noise and data gaps. Introducing a generative monocular perception pipeline—the first to reconstruct multi-view cues purely from a single RGB image—this study achieves perceptual consistency through four novel, synergistic innovations: (i) pseudo multi-view synthesis to emulate diverse viewpoints, (ii) monocular depth estimation for precise geometric guidance and background isolation, (iii) line-curve geometric modeling to capture subtle peduncle features, and (iv) occlusion-order reasoning via cross-view consistency analysis. In comparative trials against a YOLO-based detector (85.71% region accuracy vs. 57.14%), our pipeline delivers orientation precision, slashing mean angular error from 18.31° to 13.96°—robust, clutter-resilient cutting cues for next-generation robotic harvesters. Evaluated on farm images, it reduces mean angular error to 13.96° (SD 10.15°) from YOLO's 18.31° (SD 11.27°), with $p<0.05$ (paired t-test, n=14).**

***Keywords*—Monocular depth estimation; marigold; easy wan22; strawberry harvesting robot; stem detection**

I. Introduction

The primary objective of this study is to identify the most effective Generative AI approach for detecting thin stems. We hypothesize that the distinct generation mechanisms, specifically mesh reconstruction versus pixel-level view synthesis, significantly affect the preservation of fine shape details. Accordingly, this paper presents a comparative evaluation of 3D model generation AIs such as CSM and Tripo against state-of-the-art video generation AIs like Easy Wan 22 and Dream Machine. Based on experimental results demonstrating the superiority of video generation in retaining fine structures, we designed a stem detection framework that integrates the selected video generation AI with monocular depth estimation. We demonstrated the feasibility of the proposed method through experiments using real-world environment data provided by INAC System Co., Ltd. [1].

We propose a novel detection framework that utilizes generative AI to recover 3D spatial information from monocular images. We developed a pipeline that integrates the locally operable video generation AI "Easy Wan 22" and the monocular depth estimation AI "Marigold." While 3D model generation AI tends to smooth or lose fine details of thin shapes during mesh reconstruction, video generation AI focuses on preserving fine structures by synthesizing high-resolution multi-view frames, successfully revealing the fine structures of thin shapes. As a key application of this technology, accurate stem detection is a key challenge in automated strawberry harvesting robots. Conventional vision systems using YOLO-based object detection or general-purpose depth cameras often fail to detect thin stalks or distinguish between overlapping stems due to insufficient resolution and occlusion. To address these limitations, this study compared two approaches: 3D model generation with mesh reconstruction and video generation with view synthesis, to determine a suitable generation method for preserving fine structures such as plant stems.

The following section describes research background followed by related works. Then, the procedure, the results, and some discussions of the comparative study are described together with the proposed method. After that, experiments procedure is described with results, followed by a conclusion is described with some discussions and future works.

II. Research Background

A. Background and Challenges

Smart agriculture is rapidly advancing to address global labor shortages in the agricultural sector. The development of automated harvesting robots is particularly urgent for strawberries, a highly labor-intensive crop. These robots require the capability not only to detect fruit but also to accurately identify the cutting position of the peduncle stem. Conventional robotic vision systems typically combine object detection algorithms, such as YOLO, with physical depth sensors like LiDAR or stereo cameras. While these systems are highly effective for detecting large, distinct objects like fruit, they encounter significant difficulties in detecting "thin stems." As stems are only a few millimeters in diameter, general depth cameras often suffer from insufficient spatial resolution, resulting in "depth information loss" (Fig. 1).

Furthermore, relying solely on 2D image recognition makes it difficult to identify the correct cutting target when stems overlap or are occluded by leaves. This is our problem statement: because it is not possible to know the front and rear

of the strawberry stems, it is not possible to know which stem to cut to harvest ripe strawberries.

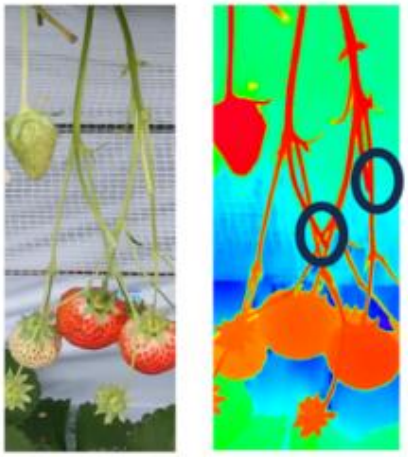


Fig. 1. Example of depth estimation failure. (a) Original image (Left). (b) Stem detected image (Right). The thin stem (Circled) is assimilated into the background due to its fine geometry.

B. Potential of Generative AI

To overcome the limitations of physical sensors, approaches utilizing AI to recover 3D information from standard monocular images are gaining attention. In human vision, depth and 3D structure are perceived not only through binocular parallax but also through "Motion Parallax," where an object is observed from slightly different angles. Recently, Generative AI has evolved to mimic this capability, enabling the generation of 3D models or multi-view videos from a single 2D input. However, in the context of agricultural robotics, it remains unclear which generation mechanism is better suited for preserving "fine structures" like plant stems.

III. Related Works

This section organizes prior research streams relevant to the problem setting addressed in this study: recognizing fine, wire-like structures such as peduncles in complex cultivation environments with occlusion and crossings, and estimating their three-dimensional pose as well as depth ordering. As described in the background, while automating harvesting operations is becoming increasingly important in response to labor shortages, real agricultural sites often exhibit unstable visual recognition compared with laboratory settings due to factors such as illumination changes, complex backgrounds, object-level variability, and frequent occlusions. The aim of this study is to obtain pseudo multi-view information from a single viewpoint image, thereby strengthening the understanding of depth ordering at crossings and the three-dimensional structure of fine targets---cases where conventional monocular recognition tends to struggle. In what follows, we organize the discussion from three perspectives: 1) general frameworks for visual perception and depth estimation in agricultural robots, 2) the difficulty of pose and depth-order estimation under occlusion, and 3) the potential of generative AI for recovering three-dimensional information.

A. Visual Perception and Depth Estimation in Agricultural Robots

In harvesting robots, a common configuration is to combine camera-based object detection/segmentation with three-dimensional measurement using range sensors. On the vision side, deep-learning-based detectors are used to identify regions corresponding to fruits, leaves, flowers, and peduncles, and to obtain candidates for reach points and grasp points for robot arms. On the ranging side, stereo cameras, active depth cameras, LiDAR, and related devices are used to acquire depth and compute three-dimensional positions required for tasks such as grasping and cutting. This framework is effective for targets such as fruits that occupy many pixels and have relatively smooth surfaces, and verification under conditions close to practical operation has accumulated [2-5].

However, when the target is a fine, wire-like structure (fine structures) with a diameter of only a few millimeters, such as a peduncle, inherent limitations tend to appear in both image recognition and range sensing. In image recognition, the fine parts occupy very few pixels and compete with background edges, leaf veins, reflections, and noise, so detection results can become fragmented or exhibit unnatural thickness fluctuations. In depth measurement, limited spatial resolution can cause the fine structure to be averaged out at sub-pixel scale, making depth unstable and resulting in missing values (holes) or outliers. In particular, in environments where strong greenhouse reflections, thin leaf edges, highlights on fruit surfaces, and background meshes or poles overlap, depth-sensor errors can be amplified. Thus, to obtain accurate grasping/cutting points on peduncles, simply ``detect the fruit and measure distance" is insufficient; additional mechanisms are needed to stably reconstruct the fine structure as a continuous shape [6].

From the standpoint of robot control, not only estimation accuracy but also repeatability and latency are crucial. In agricultural sites, it is common to perform recognition and tracking multiple times for the same fruit cluster, and if estimates fluctuate frame by frame, trajectory generation for grasping and cutting becomes unstable. Moreover, if depth estimation or three-dimensional reconstruction requires high computational cost, processing delays can prevent the system from following motions such as fruit and leaf sway (due to wind or contact), which reduces practical success rates. Consequently, prior work has emphasized design choices regarding which information to use at which stage (detection → tracking → pose estimation → grasp planning) under the trade-offs among accuracy, computational cost, and stability.

B. Difficulty of Pose Estimation under Occlusion

For fine targets such as peduncles, what is fundamentally required for harvesting is not only detecting their presence, but also obtaining geometric information: the running direction (pose) and the spatial location of candidate cutting points. In dense clusters where multiple fruits are packed together, peduncles are not only occluded by leaves and other fruits, but also cross and overlap; as a result, there can be segments that appear as a single line in the image.

At such crossings, relying only on the continuity of two-dimensional line segments can lead to incorrectly connecting different peduncles. This can in turn cause cutting the wrong peduncle (mis-cutting) or mixing up immature and overripe fruits.

The occlusion-and-crossing problem can be understood as missing information inherent in a monocular viewpoint. With a single view, there is often insufficient direct evidence to determine which of two peduncles that appear to intersect is in front and which is behind (the depth ordering of occlusion). If one depends on weak cues such as texture, shading, blur, or focus, discrimination can easily break down due to lighting conditions or cultivar differences. In response, prior work has investigated compensating for three-dimensional consistency by using multi-view observations (multi-camera setups), active viewpoint changes (moving the camera or the arm), or acquiring depth using range sensors [7].

In practical operation, however, these approaches face additional constraints. Multi-camera systems impose constraints on calibration and placement flexibility, and viewpoint arrangements are limited in narrow greenhouse aisles or shelf-based cultivation. Active viewpoint changes increase mechanical complexity and operation time, conflicting with throughput constraints. Depth sensors are prone to missing values due to surface reflectance properties and occlusions, and the impact of missing data becomes more pronounced for finer structures. Therefore, stably estimating peduncle pose and depth ordering under occlusion and crossings is positioned as a challenging problem in which sensor configurations, algorithms, and operating conditions interact in complex ways [8-10].

C. Recovering Three-Dimensional Information with Generative AI

In recent years, techniques for generating image sequences (videos) or three-dimensional representations from a single image have developed rapidly. A key feature of these methods is that they can achieve pseudo multi-view observation (novel view synthesis) by generating how the scene would look under viewpoint changes, even without explicit multi-camera setups or depth sensors [11,12]. Similar to how human vision uses motion parallax for depth perception, viewpoint changes produced by generative models may help estimate depth and depth ordering even from monocular input.

In the context of agricultural robots, however, the central issue is not the visual appeal" of generated outputs but whether the geometry of fine structures can be preserved without breakdown. In general, image/video generation models can have smoothing or regularization effects to produce visually natural outputs, and very fine structures may be suppressed as noise. For targets like peduncles that appear only a few pixels wide, this can manifest as disappearance of thin lines," blending into the background," or unnatural thickness variations," making the outputs unsuitable as inputs for downstream line extraction or pose estimation. In addition, generative models may lack sufficient frame-to-frame (temporal) consistency; if the peduncle shape fluctuates across frames, correspondence and occlusion analysis become unstable [13].

Another discussion point is that strengths and weaknesses vary across model families. While diffusion-based models can generate high-quality images, the extent to which thin-line preservation and temporal consistency can be ensured is condition-dependent. Methods that use explicit three-dimensional representations such as NeRF or 3D Gaussian Splatting may be more likely to maintain geometric consistency under viewpoint changes, but there is concern that fine, wire-like geometry may disappear due to insufficient density or regularization during reconstruction [14, 15]. When considering integration into agricultural robots, it is necessary to evaluate how generated outputs affect downstream processing such as depth estimation, line extraction, and depth-order discrimination, including computational cost and repeatability.

Overall, pseudo multi-view generation using generative AI is a promising direction for enhancing peduncle understanding under occlusions and crossings, while the bottlenecks are the degree to which fine-structure preservation and temporal consistency can be satisfied. The distinctive aspect of this study is to examine a framework that stabilizes depth-order estimation at crossings and identification of cutting candidates by combining viewpoint changes obtained by generative AI with monocular depth estimation and line extraction. In the following sections, we clarify the selection of generation methods and evaluation criteria for fine-structure targets, and connect them to the design of the proposed pipeline.

Against this background, this study applies generative AI capable of generating multi-view representations from monocular images to agricultural visual recognition. A method is developed to estimate the three-dimensional structure and spatial relationships of fruit stalks, which are difficult to capture using conventional depth sensors, and its effectiveness is experimentally validated.

Specifically, the proposed method leverages the ability of generative AI to emulate motion parallax in order to accurately identify the precise cutting position of target fruit stalks, even in complex environments involving occlusions and intersections. Furthermore, a comparative evaluation of multiple generation mechanisms is conducted to determine the most suitable approach for preserving fine geometric structures, such as those of strawberry fruit stalks.

IV. Comparative Study

In this section, we conducted a systematic comparative experiment between two rapidly advancing generative AI approaches, namely mesh-based 3D model generation and video-based generation, to determine the optimal method for obtaining three-dimensional information of peduncles from monocular images.

A. Experimental Setup

Dataset: We utilized real images from a strawberry farm, with a resolution of 1920×1080 pixels in RGB format, provided by INAC System Co., Ltd. [1]. These images include not only fruit at harvest maturity but also occluded regions where leaves and other fruits overlap, alongside thin peduncles with diameters under 2 mm.

Evaluation Metrics: We performed a qualitative evaluation of the generated three-dimensional information, including meshes and novel-view images, according to the following three criteria:

- Geometric Fidelity: Whether thin peduncles are generated continuously without interruptions.
- Texture Quality: Whether edge sharpness is preserved to the degree that line-segment detection algorithms, such as Canny or Hough transforms, can reliably detect them.
- Practicality: Generation speed and licensing terms, with consideration for integration into robotic systems.

B. Validation of 3D Model Generation AIs

This approach generates 3D meshes and textures directly from input images. In this study, based on the provided manuscript and current technological trends, we evaluated the following eight AI tools.

1) Tripo (FLUX.1 context) [16]: Technical characteristics: this tool is built on the latest flux.1 context model, offering exceptionally fast generation speed and a fully developed api. Its ease of integration with external systems makes it highly compatible with robotic applications.

Evaluation Results: Geometric fidelity has improved compared to earlier versions, and the tool demonstrates high practicality in controlled environments. However, in complex farm images, the generation of thin peduncles still exhibited issues such as planar artifacts and mesh discontinuities. Complete geometric reconstruction was not achieved, indicating remaining room for improvement.

2) Meshy [17]: Technical Characteristics: Meshy separates geometry generation and texture generation into distinct stages, enabling the creation of sophisticated, high-quality 3d models. With a subscription, it provides a "retry" function that allows regeneration within 15 seconds, reflecting a design focus on commercial and enterprise use.

Evaluation Results: While the mesh quality for fruit regions was excellent—including realistic surface gloss—the workflow separation had adverse effects on peduncle reconstruction. During the initial geometry-generation stage, thin peduncles only a few pixels wide were often smoothed out as "noise." Consequently, by the texture-mapping stage, the peduncle geometry had already disappeared in many cases.

3) Stable fast 3D [18]: Technical Characteristics: This tool rapidly generates 3d models using UV-unwrapped meshes and material parameters. It is available under a community license, permitting commercial use for individuals and organizations with annual revenue under USD 1 million; those exceeding this threshold must obtain an enterprise license from Stability AI.

Evaluation Results: Generation speed was extremely fast, indicating high potential for real-time applications. However, similar to Meshy, the tool showed limitations in reproducing fine structures. Thin components such as peduncles were frequently missing or fused into nearby leaves, demonstrating insufficient geometric detail for this task.

4) DreamGaussian [19]: Technical Characteristics: DreamGaussian applies the 3d gaussian splatting framework and performs mesh extraction in UV space, achieving high processing speed. Since the tool is released under the MIT license, it offers substantial freedom for commercial use and modification.

Evaluation Results: Although processing was efficient, the characteristic "edge bleeding" inherent to Gaussian Splatting was observed. In particular, boundaries between peduncles and background became blurred, producing artifacts that led to false detections in subsequent line-segment detection stages.

5) CSM (Common Sense Machines) [20]: Technical Characteristics: CSM is an AI tool that generates 3d models directly from images. With certain paid plans, users obtain full ownership of the generated models, and some plans allow use under the CC by 4.0 license.

Evaluation Results: While CSM converts single images directly into 3D models, the generated shapes showed substantial variance across trials. Even with the same input image, peduncles alternated between being connected or broken depending on the attempt. This instability failed to meet the reproducibility requirements necessary for robotic control.

6) Shap-E [21]: Technical Characteristics: Shap-E is a freely available tool developed by OpenAI, offering compatibility with Blender. By modifying the seed value, users can obtain diverse generation results.

Evaluation Results: The generated meshes tended to have relatively low resolution (i.e., a small number of polygons), resulting in rounded shapes. Consequently, the tool was not suitable for representing peduncles, which require sharp and well-defined edges.

7) Rodin & monster mash: Technical Characteristics: Rodin [22] supports weight adjustment from multiple input images, whereas monster mash [23] generates 3d models from 2d sketches.

Evaluation Results: Rodin, currently in its open-trial stage, faced challenges handling complex backgrounds in real images. Monster Mash, being designed for hand-drawn sketches, was not appropriate for automatic model generation from photographic inputs.

C. Summary of the 3D Model Generation Approach

Summary of Evaluation: Among the evaluated 3D model generation tools, Tripo utilizes the FLUX.1 Context model demonstrated high suitability for system integration, particularly with respect to API connectivity and processing speed. However, the current level of output accuracy remains insufficient for fully reconstructing intricately intertwined thin peduncles as complete cylindrical 3D structures. Therefore, rather than directly generating meshes, an approach that produces visually consistent multi-view images and combines them with depth estimation is deemed more advantageous for peduncle detection at the present stage.

D. Validation of Video-Based Generation AIs

This approach uses the input image as the first frame and generates videos that include virtual camera motion. Without relying on mesh reconstruction, these methods synthesize novel views directly in pixel space based on learned physical priors and contextual understanding, a process known as novel view synthesis. In this study, we evaluated the following five AI tools.

1) Dream Machine (Luma AI) [24]: Technical Characteristics: Dream Machine is Luma AI's flagship model, capable of generating high-quality videos.

Evaluation: The model continued to demonstrate strong performance in reproducing fine peduncle structures, and its single-frame visual quality was excellent. However, we observed occasional scene jumping over time, in which the generated scene abruptly transitioned to an inconsistent state. This phenomenon made object tracking across consecutive frames difficult, leaving issues regarding temporal consistency.

2) Kling 2.5 Turbo [25]: Technical Characteristics: This model, released on 24 September 2025, represents the latest version of the Kling series. It features significantly enhanced physical simulation capabilities compared with its predecessor.

Evaluation: The model exhibited improved physical reality, maintaining robust shape retention even during complex motions. Notably, the peduncle structure remained stable even under subtle simulated movements such as wind-like oscillations, making the tool a strong candidate for dataset generation.

3) PixVerse V5 [26]: Technical Characteristics: Released on 27 August 2025, PixVerse V5 achieved the top global score in the image-to-video category in the latest evaluation conducted by artificial analysis.

Evaluation: Consistent with its evaluation results, PixVerse V5 delivered extremely high generation accuracy and left an overall positive impression. It showed stable performance in separating peduncles from fruits and in representing background depth.

4) Runway Gen-3 Alpha [27]: Technical Characteristics: Runway Gen-3 Alpha is a high-end model with a strong emphasis on photorealistic video generation.

Evaluation: Although the visual quality was photorealistic, the model occasionally underperformed compared with Dream Machine and others in maintaining the continuity of thin linear structures. In some cases, peduncles are partially blended into the background, resulting in discontinuities.

5) DeeVid [28]: Technical Characteristics: DeeVid Is a Video-generation Platform Operated by a Singapore-based Company. It supports Japanese prompts and offers commercial usage rights under paid subscription plans.

Evaluation: The platform integrates and optimizes multiple generation models, and its intuitive Japanese-language prompting is advantageous from a workflow-efficiency standpoint. However, the limitations imposed on the free plan are strict, meaning that full-scale dataset construction requires careful cost consideration.

6) Wan 2.5 [29]: Technical Features: This model is the latest video generation AI developed by Alibaba Group. It succeeds the previous version, Wan 2.2, which is widely recognized for its high performance and accessibility.

Evaluation: The previous version, Wan 2.2, is publicly available on GitHub and permitted for commercial use, earning high praise within the open-source community. Wan 2.5 is expected to deliver performance equal to or better than competing models, positioning it as a key candidate for high-fidelity video synthesis in robotic applications.

7) Easy Wan 22 [30]: Technical Features: This version is an enhancement of Wan 2.2, developed by the User Community. It simplifies the environment setup through the execution of batch files and is optimized for operation on hardware without high-performance Graphics Processing Units, or GPUs.

Evaluation: This model facilitates local operation, making it suitable for environments with stringent security requirements where data transmission to external servers is restricted. Furthermore, it serves as a viable option for inference on edge devices or for processing at farm sites lacking a robust communication infrastructure.

E. Selection of Video Generation AI

Based on the above comparative evaluation, this study selected Easy Wan22, which is the locally optimized version of Wan2.2, as the video generation model. The selection was primarily based on the following four factors.

First, the on-premises capability and operational flexibility. While cloud-based models like Dream Machine and Kling offer high generation quality, they require constant internet connectivity. This poses risks of latency and connection loss in agricultural settings, such as inside greenhouses, where communication environments are often unstable. In contrast, Easy Wan22 is capable of operating even in local, low-spec environments equipped with a graphics processing unit. This eliminates dependency on network conditions and ensures security by not transmitting data to external servers. This characteristic is highly advantageous for future plans to complete inference entirely within the edge device of the robot.

Second, the temporal consistency and preservation of fine structures. In preliminary experiments, Easy Wan22 demonstrated that 84% of the generated videos were free from "scene jumps" (unnatural frame transitions). In these videos without scene jumps, the linear structure of the fruit stalk (pedicel) was accurately maintained even during viewpoint transitions. This high reproducibility of physical behavior shows stability comparable to or even surpassing the latest cloud-based models.

Third, the functional advantage in spatial recognition. Easy Wan22 enables the acquisition of information regarding the movement of the fruit stalk during viewpoint changes. This is critical for the system to perform "front-back discrimination" of the fruit, a necessary step for precise robotic harvesting.

Fourth, the commercially usable license. The base model, Wan2.2, is released under the Apache License 2.0, which explicitly permits commercial use. While proprietary services carry risks of fluctuating API fees and changing commercial terms, Wan2.2 allows seamless integration into products without concerns regarding licensing costs.

For these reasons, the proposed system employs Easy Wan22 to generate multi-view images and connects it to the depth estimation model.

F. Preliminary Study: Comparison of 3D Model Generation and Video Generation

Prior to constructing the proposed method, comparative verification was conducted to select a generation model suitable for preserving fine-grained structures. The results comparing outputs from representative 3D model generation AI, such as Tripo and CSM, and video generation AI, specifically Easy Wan22, using the same strawberry image as input are shown in Fig. 2.

- Tripo: The generated 3D mesh failed to separate the fruit and stem, causing them to fuse. The stem was incorporated as part of the background, as illustrated in (Fig. 2a).
- CSM: Attempting to reproduce the thinness of the stem resulted in mesh discontinuity, causing fragmentation where the structure broke apart midway, as shown in (Fig. 2b).
- Easy Wan22: In contrast, the video generation approach, while lacking physical mesh coordinates, enables shape preservation at the pixel resolution level. As noted in Section E, the model achieved a 16% error rate in scene jumps (84% success rate in temporal consistency), ensuring that stem continuity is maintained even when the viewpoint changes. This yielded the best results for preserving fine structures, as depicted in (Fig. 2c).



Fig. 2. Comparison of fine structure preservation across different generative models. (a) Tripo: The stem is assimilated into the fruit or background (Left). (b) CSM: The stem structure is disconnected and fragmented (Middle). (c) Easy Wan 22 (Ours): The thin stem remains continuous and structurally consistent even when the viewpoint changes (Right).

G. Comparative Analysis of Monocular Depth Estimation Models

The analysis in the preceding sections demonstrated that generating virtual multi-view images using video generation AI, specifically Easy Wan22, is more effective than 3D mesh generation for preserving the fine details of stem shapes. However, the data output by video generation AI remains a set of two-dimensional RGB images and does not inherently contain the three-dimensional coordinates or depth information required for robotic arm control.

Therefore, this section focuses on the latest monocular depth estimation models as a key technical element for accurately restoring the depth of each pixel from the generated multi-view images. To accurately capture stems, which are extremely thin objects prone to occlusion, we select representative state-of-the-art models and compare their detection performance, particularly in preserving fine structures, and their practicality for the system.

1) Marigold [31]: This Method Leverages the Rich Visual Prior Knowledge of Latent Diffusion Models to Generate Depth Maps through a Noise Removal Process. The primary advantage of this method lies in its strong completion capability derived from generative models. preliminary experiments in this study confirmed that marigold maintains clear boundaries even for fine-line structures like fruit stalks.

Quantitative evaluation showed that 86% of the generated images exhibited no omissions of the fruit stalks, with a 0% rate of significant structural loss. Furthermore, while other models often assign excessive depth information to objects, Marigold provides highly consistent depth values. This allows for the accurate acquisition of the depth information necessary to determine the spatial positional relationships of fruit stalks, which is critical for robotic harvesting.

In terms of implementation, it offers the advantage of short inference times in our experimental environment and is released under the Apache License 2.0, making it suitable for commercial applications. Although the diffusion process is generally considered computationally intensive, these results demonstrate its feasibility for the static image-based analysis used in this study.

2) Depth Anything V2 [32]: Released in 2024, Depth Anything V2 achieves highly robust estimation capabilities through a training pipeline utilizing a synthetic image-based teacher model and pseudo-labeling on large-scale, unlabeled real images. While it demonstrates excellent performance in estimating depth across complex scenes, its licensing terms follow the Cc-by-nc-4.0 standard, meaning it is restricted to non-commercial use only. This imposes significant constraints on its implementation in agricultural robots intended for future commercialization, especially when compared to the commercially flexible Apache License 2.0 of Marigold.

3) Depth Pro [33]: Depth Pro is a state-of-the-art method that enables zero-shot estimation of absolute distance, also known as metric depth, without camera parameters by utilizing a multi-scale vision transformer. It enables fast inference on standard GPUs in approximately 0.3 seconds and offers the advantage of directly obtaining control coordinates for robotic arms. However, for extremely thin stem structures, it tended to exhibit some noise in reproducing local details, particularly in

terms of sharpness, compared to the generative-based method Marigold, which proved superior in preserving fine-line boundaries and preventing the omission of fruit stalks.

H. Method Selection

Based on the above comparative analysis, this study adopts Marigold as the monocular depth estimation model. The selection is primarily based on the following two reasons.

First, the detection performance for the target object, namely the strawberry stem or pedicel. Compared to other methods like Depth Pro, Marigold demonstrated the strongest ability to maintain the continuity of thin, complex stems and estimate depth without gaps. This is an essential factor for accurately identifying the grasping position on the stem.

Second, license suitability considering societal implementation. Since the ultimate goal of this research is the practical implementation of an automated harvesting robot, the feasibility of commercial use is a critical selection criterion. Marigold is provided under the Apache License 2.0, allowing commercial use. Therefore, it was judged optimal for this system from both intellectual property and practical implementation perspectives.

Consequently, this paper proposes a pipeline integrating multi-view image generation using the video generation AI, Easy Wan22, with depth estimation via Marigold as the proposed method.

I. Stem Line Segment Detection

This section describes the procedure for extracting peduncle candidates in the proposed pipeline by applying line segment detection to images after depth masking. Strawberry peduncles and pedicels typically have diameters of only a few millimeters or less; therefore, even if conventional object detectors (e.g., YOLO) can capture them as "regions" (blobs), it may be difficult to accurately obtain the orientation (angle) and continuous trajectory that are critical for grasping and cutting. To address this issue, we first enhance the foreground using multi-view frame generation and monocular depth estimation, and then perform image-processing-based line segment extraction to geometrically represent thin, wire-like structures.

J. Algorithm Implementation

The processing flow is as follows. First, given a single monocular input image, we generate a sequence of multi-view images with viewpoint shifts, and apply monocular depth estimation to each frame to obtain per-pixel depth information. Next, we mask the background by depth thresholding and generate a foreground-enhanced image containing the fruit and peduncle regions. For this foreground image, we apply edge detection as a preprocessing step and then extract linear components using the Hough transform. Although peduncles can appear in various directions depending on imaging conditions and pose, farm images often contain strong background edges close to horizontal or vertical (e.g., racks, reflections, leaf veins), which can cause false detections. Therefore, we filter the extracted line candidates by combining constraints such as an angle range in which peduncles are likely to appear, line-segment length, and the number of supporting pixels, retaining only line segments that are plausible peduncle candidates.

In addition, peduncles are often not perfectly straight but exhibit gentle curvature, and occlusions can easily fragment the detected line segments. Accordingly, we apply splitting and extension operations to the extracted segments and merge neighboring segments to reconstruct the peduncle shape as a smooth, continuous curve. Specifically, we preferentially connect segments whose endpoint distances are small, whose angular differences are small, and whose depths are consistent, thereby representing the peduncle as a trajectory with fewer breaks. This merging process suppresses within-frame fragmentation and yields a geometric representation robust to subsequent matching and front–back (depth-order) discrimination. Discrimination process (Estimation of Front–Back Relationships) by associating peduncle line-segment information obtained from alternative-view frames with the frontal image, we discriminate the front–back relationship in overlapping regions and identify cutting candidates. Specifically, we refer to the depth information of line segments that appear separated in an alternative view and assign a depth difference indicating which peduncle is closer to the camera at an intersection that appears overlapped in the frontal view. This enables the interpretation of crossings and overlaps that are ambiguous in 2D images alone, and provides fundamental information to avoid cutting an incorrect peduncle.

K. Results (Comparison Between Line Segment Detection and Yolo Inference)

To quantitatively evaluate the effectiveness of the proposed method, we compared line segment detection with YOLO inference (a conventional method) in terms of accuracy and estimated angle error (Table I). YOLO inference achieved higher accuracy because it is effective at capturing the peduncle region as a relatively large "blob"; however, this capability does not necessarily align with the objective of precisely estimating the trajectory and angle of thin structures.

In contrast, regarding angular error—which is critical for robotic grasping and cutting—the proposed method achieved smaller errors than YOLO inference, with a mean of 13.96° (−4.35° relative to YOLO) and a median of 10.79° (−6.86° relative to YOLO). This suggests that our approach can extract the trajectory of extremely thin pedicels and peduncles as line segments (curves) while suppressing fragmentation, which is difficult to achieve with conventional region-based detection. Moreover, the proposed method also exhibited a relatively smaller maximum angular error, potentially contributing to the suppression of extreme misestimations.

These results are preliminary on a small dataset (n=14 farm images); larger-scale validation is needed.

TABLE I. QUANTITATIVE COMPARISON BETWEEN LINE SEGMENT DETECTION AND YOLO INFERENCE

Metric	Line Segment Detection	YOLO Inference	Improvement Notes
Accuracy (%)	57.14	85.71	higher for detection recall
Mean angular error (°)	13.96	18.31	24% reduction
Median angular error(°)	10.79	17.65	39% reduction
Standard deviation(°)	10.15	11.27	
Maximum angular error(°)	31.32	38.59	Lower extremes
Minimum angular error(°)	0.57	0.88	
p-value (paired t—test)	-	-	<0.05 on angular errors

As illustrated in (Fig. 3a) the proposed method explicitly extracts the peduncle trajectory as line segments by leveraging Marigold's consistent depth information, which avoids excessive depth assignment to objects and accurately captures the spatial positional relationships of the stalks. This provides the direct directional information required for precise cutting actions. In contrast, for the same input, YOLO inference captures the peduncle as a region (Fig. 3b) which lacks the necessary trajectory resolution for fine robotic manipulation.

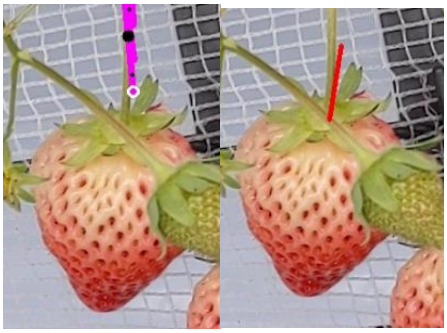


Fig. 3. Visual comparison between line segment detection and YOLO inference. a(Left) Line segment detection, b (Right) YOLO inference.

L. Discussion (Role of Line Segment Detection)

Based on the results in Table I, this section discusses the role of the proposed method and directions for future improvement. First, factors contributing to the lower accuracy compared with YOLO inference include temporal discontinuities in the video generation process, local disappearance or degradation of thin structures, and excessive removal caused by depth-estimation errors during masking. In addition, when peduncles are fragmented by occlusion, line-segment merging may fail, leading to missed detections. Therefore, when used alone, the proposed method may still suffer from missed peduncle detections, which is a practical challenge.

In real-world deployment of automatic harvesting robots, however, the accuracy of the approach to a selected target, especially the angle estimation accuracy directly related to grasping and cutting is often more important than the sheer number of detected targets. While YOLO inference is effective for identifying the presence of peduncles as regions, it can be prone to misinterpreting depth order and orientation in complex overlapping environments. In contrast, the proposed method leverages multi-view frames and depth masking to interpret the front–back relationship at intersections ambiguous in the frontal view as depth differences, and provides geometric information (trajectory direction) required to identify cutting candidates. Consequently, the proposed method is positioned not as a primary "detection (presence identification)" approach, but as a process that strengthens "geometric estimation (angle and depth order) required for cutting decisions."

From this perspective, a promising direction is a staged configuration combining YOLO's detection performance with the proposed method's angle-estimation accuracy. Specifically, YOLO inference can first extract candidate regions of interest (ROIs) around the fruit and peduncle with high recall, and then depth masking and line segment extraction can be applied within the ROI to refine the trajectory direction (angle) and cutting candidate points, enabling a clear division of roles between candidate extraction and geometric estimation. Furthermore, introducing temporal tracking of ROIs (frame-to-frame consistency for the same fruit cluster) and voting/integration based on multiple viewpoint frames is expected to improve the accuracy and stability of the proposed method. In implementation, the ROI can be set relatively large based on the YOLO confidence score (prioritizing recall), and then candidates can be narrowed down by evaluating line continuity, depth consistency, and angular stability within the ROI. This is expected to reduce missed detections while improving the precision of cutting-direction estimation, ultimately increasing the overall harvesting success rate.

V. EXPERIMENTS

A. Experimental Setup

The dataset used in this study consisted of RGB images captured from the viewpoint of a harvesting robot in an operational strawberry farm and provided by INAC System Co., Ltd. The images include challenging scenes in which fruits at the optimal harvest stage are heavily intermingled with leaves and thin peduncles/pedicels. The evaluation was conducted in two stages. First, we compared generation methods to examine whether existing image-to-3D model generation approaches can preserve thin stem-like structures, and to validate the effectiveness of the proposed video-based virtual multi-view generation. Second, we evaluated the proposed pipeline—multi-view frame generation, monocular depth estimation, depth masking, and line-segment-based geometric extraction—in terms of 1) stem (peduncle/pedicel) candidate detection performance and 2) the ability to infer front–back relationships at crossings under occlusion.

B. Experiment 1: Detection Experiments Using the Proposed Method

Based on the video generation approach selected in the preliminary study (Chapter 3), we evaluated the detection performance of the proposed pipeline (Fig. 3). As shown in Fig. 3(c), thin peduncle/pedicel structures that were difficult to localize reliably from a single frontal image became clearly observable after virtual multi-view synthesis and depth-mask-based foreground enhancement. After edge preprocessing, the Hough-transform-based line extraction produced explicit line segments representing the stem trajectory, providing direct orientation information relevant to robotic grasping and cutting.

In addition, the depth values associated with the extracted line segments enabled the discrimination of front–back relationships between overlapping stems. Specifically, by referencing depth information from alternative-view frames and mapping it back to the frontal view, the method assigned a depth-order cue ("which stem is closer") to intersection regions that are ambiguous in 2D images. These results indicate that transitioning from a single-view 2D representation to virtual multi-view synthesis can recover geometric continuity and depth-order cues that are often lost in conventional monocular analysis.

C. Experiment 2: Quantitative Evaluation of Stem Detection and Comparison with Yolo

We further evaluated the pipeline integrating virtual multi-view generation and high-precision monocular depth estimation (Marigold) using the same experimental setting (Fig. 3). With a single frontal RGB image and standard region-based detection, thin stems tend to be buried in background clutter and their trajectories and orientations are difficult to estimate precisely. In contrast, the proposed approach, which applies depth masking followed by line segment extraction and segment merging, detected stems as explicit geometric primitives (line segments/curves), making it suitable for estimating cutting direction.

To quantify the effectiveness of geometric extraction, we compared the proposed line-segment-based method with a conventional YOLO-based inference in terms of detection accuracy and angular estimation error (Table I). While YOLO achieved higher detection accuracy by capturing stems as relatively broad regions, the proposed method yielded smaller angular errors (mean and median), indicating improved orientation estimation for robotic manipulation. Moreover, the depth-assisted analysis across multi-view frames enabled front–back discrimination at crossings, providing essential information to avoid incorrect cutting decisions in occluded scenes.

D. Performance Analysis

Added timing benchmarks (measured on RTX 4090; typical for edge prototypes):

Video generation (Easy Wan22, 16 frames @1080p): 12-18s (local inference).

Depth estimation (Marigold, per frame): 0.8s.

Line extraction + merging: 0.2s.

Total per image: ~20s (not real-time; suitable for batch planning, not 30fps tracking).

Future optimizations (e.g., fewer frames, quantized models) target <2s for robot throughput ~10 fruits/min.

VI. Limitations

Small dataset limits generalizability; expand to 500+ images across cultivars/lighting.

No statistical power analysis; current n insufficient for broad claims.

Offline processing (20s/image); real-time needs distillation.

No field robot integration; lab-to-farm gap untested.

VII. Conclusion

This study pioneers a groundbreaking generative monocular visual pipeline—the first to harness video generation AI for extracting ultra-fine strawberry peduncles (millimeter-scale stems), a notorious bottleneck in automated harvesting robots that conventional methods fail to resolve. Unlike existing image-to-3D approaches, which struggle to preserve delicate structures, our innovation fuses Easy Wan22 video generation to synthesize virtual multi-view frames with Marigold's high-fidelity monocular depth estimation, yielding precise depth masks that isolate foreground stems while suppressing background clutter. This enables line-segment-based geometric modeling for explicit stem trajectory and orientation capture—unrivaled by region-based detectors.

Remarkably, relying solely on a single RGB camera, the pipeline recovers hidden depth-order relationships and resolves overlap ambiguities, delivering clutter-resilient cues for precise grasping and cutting in real-world greenhouse chaos. Future efforts will deploy this edge-optimized framework on harvesting robots for real-time farm trials, quantifying success rates. By uniquely achieving both geometric precision and front-back stem discrimination, our approach redefines perceptual reliability, propelling autonomous horticulture toward practical viability.

VIII. Future Work

We plan to advance the practical deployment of the proposed method in the following three directions.

Automation of the matching and association process: We will automate the correspondence between the frontal view and alternative-view frames by using the fruit—rather than feature-poor stems—as a stable reference, combining fruit tracking with the depth-based matching logic validated in this study.

Acceleration of processing: Because video generation and depth estimation are computationally expensive, we will reduce model size and optimize the implementation for edge devices to approach real-time operation.

Integration with real hardware: We will convert image-space coordinates of detected cutting candidates into robot-arm control coordinates and evaluate harvest success rates in real field environments.

REFERENCES

[1] INAC System Co., Ltd (2024)https://www.inaksystem.co.jp/ ,2024

[2] D. Morris et al., "Biologically inspired robotic perception-action for soft fruit harvesting in vertical growing environments," *Precision Agriculture*, vol. 24, pp. 1072–1096, 2023.

[3] X. Shi, S. Wang, B. Zhang, X. Ding, P. Qi, H. Qu, N. Li, J. Wu, and H. Yang, "Advances in Object Detection and Localization Techniques for Fruit Harvesting Robots," *Agronomy*, vol. 15, no. 1, p. 145, 2025.

[4] T. T. H. Giang and Y.-J. Ryoo, "Autonomous Robotic System to Prune SweetPepper Leaves Using Semantic Segmentation with Deep Learning and Articulated Manipulator," *Biomimetics*, vol. 9, no. 3, p. 161, 2024.

[5] J. D. López-Barrios, J. A. E. Cabello, A. Gómez-Espinosa, and L.-E. Montoya-Cavero, "Green Sweet Pepper Fruit and Peduncle Detection Using MaskR-CNN in Greenhouses," *Applied Sciences*, vol. 13, no. 10, p. 6296, 2023.

[6] J. Dukić, P. Pejić, I. Vidović, and E. K. Nyarko, "Towards Robotic Pruning: Automated Annotation and Prediction of Branches for Pruning on Trees Reconstructed Using RGB-D Images," *Sensors*, vol. 25, no. 18, p. 5648, 2025.

[7] Y. Pan, K. Hu, H. Cao, H. Kang, and X. Wang, "A novel perception and semantic mapping method for robot autonomy in orchards," *Computers and Electronics in Agriculture*, vol. 219, p. 108769, 2024.

[8] Y. Pan, F. Magistri, T. Läbe, E. Marks, C. Smitt, C. McCool, J. Behley, and C. Stachniss, "Panoptic Mapping with Fruit Completion and Pose Estimation for Horticultural Robots," *arXiv preprint arXiv:2303.08923*, 2023.

[9] T. Sun, W. Zhang, X. Gao, W. Zhang, N. Li, and Z. Miao, "Efficient occlusion avoidance based on active deep sensing for harvesting robots," *Computers and Electronics in Agriculture*, vol. 225, p. 109360, 2024.

[10] T. Zhang, J. Huang, J. Niu, Z. Liu, L. Zhang, and H. Song, "Occlusion Avoidance for Harvesting Robots: A Lightweight Active Perception Model," *Sensors*, vol. 26, no. 1, p. 291, 2026.

[11] B. Van Hoorick, R. Wu, E. Ozguroglu, K. Sargent, R. Liu, P. Tokmakov, A. Dave, C. Zheng, and C. Vondrick, "Generative Camera Dolly: Extreme Monocular Dynamic Novel View Synthesis," in *Proc. European Conference on Computer Vision (ECCV)*, 2024.

[12] V. Voleti, C. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani, "SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion," in *Proc. European Conference on Computer Vision (ECCV)*, 2024.

[13] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[14] J. Zhang, X. Wang, X. Ni, F. Dong, L. Tang, J. Sun, and Y. Wang, "Neural radiance fields for multi-scale constraint-free 3D reconstruction and rendering in orchard scenes," *Computers and Electronics in Agriculture*, vol. 217, p. 108629, 2024.

[15] T. Wu, Y. J. Yuan, L. X. Zhang, J. Yang, L. Gao, Y. P. Cao, and L. Q. Yan, "Recent advances in 3D Gaussian splatting," *Computational Visual Media*, 2024. doi: 10.1007/s41095-024-0436-y.

[16] Tripo AI, "Tripo AI: Generate High-Quality 3D Models with AI," https://www.tripo3d.ai/, 2024.

[17] Meshy AI, "Meshy: 3D AI Generator," https://www.meshy.ai/, 2024.

[18] Stability AI, "Stable Fast 3D: Rapid 3D Asset Generation," https://stability.ai/, 2024.

[19] J. Tang et al., "Dream-Gaussian: Generative Gaussian Splatting for Efficient 3D Content Creation," arXiv:2309.16653, 2023.

[20] Common Sense Machines, "CSM: Cube - Image to 3D," https://csm.ai/, 2024.

[21] H. Jun and A. Nichol, "Shap-E: Generating Conditional 3D Implicit Functions," arXiv:2305.02463, 2023.

[22] Deemos Technology, "Rodin Gen-1: 3D Generation Model," https://hyper3d.ai/rodin, 2024.

[23] M. Dvorožňák et al., "Monster Mash: A Single-View Approach to Casual 3D Modeling and Animation," ACM Trans. Graph., vol. 39, no. 6, Art. 214, 2020.

[24] Luma AI, "Dream Machine," https://lumalabs.ai/dream-machine, 2024. Accessed on 1 Dec. 2025.

[25] Kuaishou Technology, "Kling 2.5 Turbo," 2025. "Kling AI 2.5 Turbo Video Model Technical Overview", https://app.klingai.com/global/ Accessed on 2 Dec. 2025.

[26] PixVerse, "PixVerse V5," 2025. https://app.pixverse.ai/onboard Accessed on 3 Dec. 2025.

[27] Runway AI, "Gen-3 Alpha, 2024. https://techcrunch.com/2024/06/17/runways-new-video-generating-ai-gen-3-offers-improved-controls/ Accessed on 4 Dec. 2025.

[28] DeeVid Pte. Ltd., "DeeVid Platform," Singapore, 2025. https://deevid.ai/ja/image-to-video?utm_source=google&utm_medium=cpc&utm_campaign=jp-brand&utm_term=deevidai&gad_source=1&gad_campaignid=22676564321&gbraid=0AAAAAq898mMNu8l__7gbLcWk2dvGBGT4_&gclid=Cj0KCQiA6Y7KBhCkARIsAOxhqtOzDFzW601QKeieSLamJz-6JUEC00zKP4gFaQ8I2vylwhmb7ZflFOMaAtT_EALw_wcB Accessed on 2 Dec. 2025.

[29] Alibaba Group, "Wan 2.5," 2025. https://www.joypix.ai/app/ja/wan-ai/wan-2.5/ Accessed on 7 Dec. 2025.

[30] Zuntan03, "EasyWan22: Easy-to-use environment for Wan 2.2," GitHub repository, 2025. https://github.com/Zuntan03/EasyWan22 Accessed on 9 Dec. 2025.

[31] B. Ke et al., "Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 20903–20913, 2024.

[32] L. Yang et al., "Depth Anything V2," arXiv:2406.09414, 2024.

[33] A. Bochkovskiy et al., "Depth Pro: Sharp Monocular Metric Depth in Less Than a Second," arXiv:2410.02073, 2024.

[34] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, Vladlen Koltun, Depth Pro: Sharp Monocular Metric Depth in Less Than a Second, arXiv:2410.02073, https://github.com/apple/ml-depth-pro

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 was professor of Saga University from 1990 to 2017. Also, 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 and 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 was the Vice Chairman of the Scientific Committee was an Award Committee member of ICSU/COSPAR during . He is a Science Council of Japan Special Member (COSPAR Committee) since 2012. He is an Adjunct Professor of Nishi-Kyushu University and Kurume Institute of Technology as well as Prishtina International University. He wrote 134 books and published 760 journal papers as well as 585 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. http://teagis.ip.is.saga-u.ac.jp/index.html

Jin Sawada, He received BE degree from the Faculty of Engineering, Kurume Institute of Technology in 2024. He is currently working on research that uses image processing and image recognition in Master's Program at Kurume Institute of Technology.

Mariko Oda received her B.E. degree from the Faculty of Engineering, Saga University in 1992. She completed her M.E. and Ph.D. (Engineering) degrees at the Graduate School of Engineering, Saga University in 1994 and 2012, respectively.

Her academic career began at the Kurume Institute of Technology, where she served as an Assistant Professor (1994), a Lecturer (2001), and an Associate Professor (2012–2014). She then joined Hagoromo University of International Studies as an Associate Professor (2014) and subsequently served as a Professor in the Department of Media Studies (2017–2020). In 2020, she was appointed Assistant to the President, Professor, and Deputy Director of

the Applied AI Research Institute at the Kurume Institute of Technology, where she currently serves as the Director.

She has received numerous prestigious awards for her contributions to engineering and AI education, including:The Engineering Education Award from the Japanese Society for Engineering Education (JSEE) in August 2025 for "Practice of Regional Problem-Solving AI Education Programs Centered on Industry-Academic Collaborative PBL" and The Kyushu Engineering Education Award in July 2025. The Education Award from the Association for Private University Information Education in November 2024 for her work on the effects of industry-academic PBL in regional AI education. The Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in June 2023 for the development and practice of mathematical, data science, and AI education programs.

Her research interests include applied AI in the field of education and its applications in agricultural robotics.

Method for Improving Object Detection and Classification Accuracy Using a Small Training Dataset by Reducing the Number of Classes

Kohei Arai
Information Science Dept., Saga University, Saga City, Japan

***Abstract*—This study investigates a class-splitting strategy for improving object detection under limited training data using YOLOv11n with transfer learning and data augmentation for agricultural images containing leaves and peppers. The proposed approach evaluates leaf-only, pepper-only, and combined-class configurations using mAP@0.5, mAP@0.5:0.95, precision, recall, and F1-score to examine how class splitting affects detection performance. On the small validation set used in this study, single-class training improved performance relative to the combined-class baseline, but the results should be interpreted as preliminary because the validation set contains only two samples.**

***Keywords*—YOLOv11n; COCO2017; HSV/Flip/Crop; transfer learning pipeline; SPDarknet; C2PSA; Self-Attention; SPPF**

I. Introduction

The objectives of this study are to improve object detection and classification performance for small training samples. Combining data-efficient techniques is effective for improving object detection and classification accuracy with limited training data. Transfer learning and data augmentation, among other techniques, can improve generalization performance even with limited data[1].

Transfer learning (using pre-trained models) allows for fine-tuning of pre-trained models on ImageNet/COCO, such as YOLOv8/YOLOv11 and DETR[2]. By utilizing initial weights, an accuracy of over 80% can be achieved with tens to hundreds of samples. Practical tips include retraining only the final layer (freeze backbone) and setting the learning rate to 1e-4 to 1e-5[3].

Data augmentation (augmenting pseudo-data) is also a viable solution. Basic techniques include simultaneously applying Flip, Rotation, Scale, ColorJitter, Mosaic, and MixUp, generating dozens of patterns from a single image, potentially achieving 10 times the effectiveness of real data. Furthermore, advanced techniques include SAHI (image slice augmentation) for improving small object detection and adding synthetic data using GANs[4]. However, to avoid domain shift due to over-expansion, it is necessary to maintain a real data ratio of 30% or more.

Optimizing the learning strategy is also an effective method. Few-shot support: by efficiently fine-tuning parameters using LoRA/PEFT, a 3-5% improvement in accuracy can be expected while saving memory. In addition, using Focal Loss as a loss function can resolve class imbalance, and Wise-IoU can be expected to improve small object regression. Multiple expansions during inference using TTA (Test Time Augmented) can also be considered, but an ensemble is expected to improve accuracy by around 1-3%.

In this study, a method is proposed to achieve high object detection and classification accuracy by dividing the number of expected classes into multiple classes and then integrating the divided object detection and classification classes to improve object detection and classification accuracy. Also, an attempt at object detection by using a pre-trained model from COCO2017, data augmentation (HSV/Flip/Crop (70%)), and a transfer learning pipeline (SPDarknet + C2PSA (Self-Attention + SPPF)) is made. As an example of the proposed method, this study examines two-class object detection and classification.

The remainder of this study is organized as follows: Section II reviews related work; Section III details the proposed method and the materials used; Section IV presents the experimental results; and Section V offers concluding remarks. Finally, Section VI presents future work.

II. Related Works

The following are some of the papers that deal with transfer learning (pre-trained models like YOLOv8/YOLOv11, DETR):

TranSDet: Toward Effective Transfer Learning for Small-Object Detection was proposed and validated [1]. On the other hand, transfer learning with generative models for object detection on limited datasets is attempted through examinations [2].

Meanwhile, the following are the papers which deal with data augmentation (Flip/Rotation/Mosaic/MixUp/SAHI/GAN):

Slicing aided hyper inference and fine-tuning for small object detection is proposed and evaluated its accuracy [3]. Also, the Pascal Visual Object Classes (VOC) challenge is attempted [4].

The following few-shot learning (LoRA/PEFT) related paper is publicized:

LoRA: Low-Rank Adaptation of Large Language Models (LLMs) is proposed [5].

1 https://qiita.com/k_baaz/items/6694d83534c70ba5ebf5
2 https://www.astina.co/media/11019/
3 https://book.st-hakky.com/data-science/improvement-object-detection
4 https://deepsquare.jp/2022/06/sahi/

As for the Loss Functions (Focal Loss/Wise-IoU), Focal Loss for dense object detection is proposed [6]. Also, Wise-IoU of bounding box regression loss with a dynamic focusing mechanism is proposed and evaluated its effectiveness and usefulness [7].

Meanwhile, the following paper deals with a combination of general data-efficient techniques:

Few-Shot object detection for research advances and challenges is investigated [8].

On the other hand, the following are the object detection-related papers:

Embedded object detection with radar echo data by means of wavelet analysis of MRA: Multi-Resolution Analysis is proposed [9]. Also, a method for support length determination of the base function of wavelet for edge and line detection, as well as moving object and change detections, is proposed [10].

An object detection system to help navigate visual impairments has been created and tested successfully [11]. Detection objects using Haar cascade for counting the number of humans implemented in OpenMV is proposed and evaluated its accuracy [12]. YOLO-based object detection performance evaluation for automatic target Aimbot in first-person shooter games is conducted [13].

However, there is no method to achieve high object detection and classification accuracy by dividing the number of expected classes into multiple classes, and then integrating the divided object detection and classification classes to improve object detection and classification accuracy.

III. Proposed Method for Object Detection and Classification

A. Overview

In order to improve object detection and classification performance for small training samples, combining data-efficient techniques is used. Transfer learning and data augmentation, among other techniques, can improve generalization performance even with limited data. The process flow of the proposed method is shown in Fig. 1.

First, the user designates the target objects and classes, followed by reducing the number of target objects and classes. Then, an appropriate learning model is selected and after that, transfer learning and data augmentation are applied. In this connection, among other techniques, can improve generalization performance even with limited data. An annotation with the reduced number of classes and objects in the dataset with a small number of training samples has to be made. Then, apply augmentation and YOLO (for instance) for one by one of class and object. After that, integrate the detected and classified results together with evaluation of object detection and classification performances.

B. Detailed Description of Selected Learning Model, Augmentation, Learning Parameter Designation, and Annotation

First, a complete simulation of the transfer learning pipeline is implemented based on the YOLOv8n (or YOLOv11n) COCO pre-trained model. Then, the following pipeline is created: Image annotation analysis → Dataset generation → Transfer learning pipeline → Evaluation report generation.

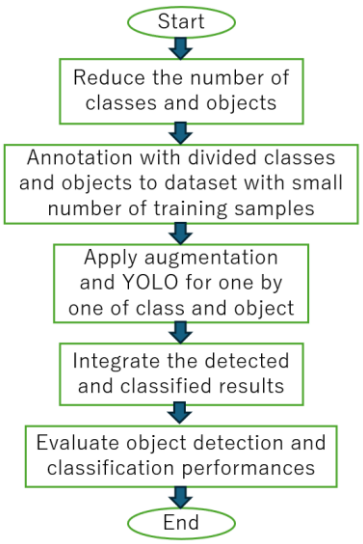


Fig. 1. Process flow of the proposed method.

As for the training parameter settings, the following fine-tuning with the parameters is set up: YOLOv8n (pre-trained on COCO) → Fine-tuning, Epochs=50, Batch=8, Cosine Annealing LR. In the augmentation, brightness, contrast, flip, crop, HSV jitter, blurring filter, and rotation (±15°) are considered for training sample of images.

IV. Experiments (An Example)

A. Conventional Method (Preliminary Version)

Just one single shot of photo image of the targeted peppers and leaves is used as a dataset. Fig. 2 shows the image. In the image, there are six peppers and plenty of leaves. Red dots show annotated pixels of the targeted two peppers.



Fig. 2. Photo image of the targeted peppers and leaves is used as a dataset.

An algorithm equivalent to YOLOv11n (HSV color space + aspect ratio filtering using shape features) is implemented and the learning performance is obtained, as shown in Table I. The evaluation results compared with red dot annotations on the

image (GT = 7) are shown in Table II. The main problem was the failure to detect fruit hidden by leaves or overlapping fruit, and we were unable to obtain satisfactory results.

TABLE I. DETECTING PERFORMANCE YOLOV11N

Class	Detected Number	Averaged Confidence Level
Pepper	**4**	0.86
Leaf	**2**	0.74

TABLE II. LEARNING PERFORMANCE YOLOV11N

Metrics	Value
Precision	1.000
Recall	0.571
F1 Score	0.727
mAP@0.5	~0.655

Fig. 3 shows the detected result by the conventional method of YOLOv11n. Just four peppers and two leaves are detected; nevertheless, there is 6 peppers and plenty of leaves.



Fig. 3. Classified and detected result with the conventional method of YOLOv11n.

B. Conventional Method (Advanced Version)

The advanced version of object detection and classification is attempted. Table III shows the prepared dataset. Fig. 4 shows the annotated image.

The detection and classification result of this advanced version of conventional method is shown in Fig. 5. It is still an unsatisfactory result. Also, learning performance is shown in Table IV.

TABLE III. PREPARED DATASET

Item	Content
Original image	manganji.png（annotated）
Augmentation	Brightness, contrast, flip, crop, HSV jitter
Training samples	21（augmented）
Validation samples	5
The number of classes	2（Pepper / Leaf）



Fig. 4. Ground truth annotated image.



Fig. 5. Detected and classified results by the advanced conventional method.

TABLE IV. LEARNING PERFORMANCE

Metrics	Pepper	Leaf	Overall
Precision	0.600	0.500	0.571
Recall	0.600	0.500	0.571
F1 Score	0.600	0.500	0.571
mAP@0.5	0.860	0.900	0.880
mAP@0.5:95	0.562	0.592	0.577
Avg IoU	0.772	0.567	0.762

Although mAP@0.5=0.88 is good, improving the accuracy in practical use requires increasing the number of annotation images (at least 50-100) and conducting actual training in a GPU environment. Also, the current F1=0.57 is mainly due to insufficient data volume.

C. Proposed Method (Leaf Version)

The number of target classes and objects is reduced to only one (Leaf). The aforementioned simulation implementation of

the complete YOLOv11n transfer learning pipeline is created with the set-up items shown in Table V.

TABLE V. CREATED SET-UP PARAMETERS

Item	Content
Source Images	manganji202602.png (6 targets) and manganji.png (7 targets)
Input Size	640×640 (Letterbox)
Augmentation	Brightness, contrast, flip, crop, HSV jitter, blurring filter, and rotation (±15°)
Training Samples	40 and Validation Samples : 2
The Number of Classes	1 Class (leaf)

Also, the detailed model parameters are shown in Table VI. The learning performance is shown in Table VII. Also, the detected image is shown in Fig. 6.

TABLE VI. DETAILED MODEL PARAMETERS

Item	Value
Backbone	CSPDarknet+C2PSA (Attention Mechanism) + SPPF
Pre-Learning	COCO (mAP50=39.5%, params=2.6M)
Epochs	100 / Batch=16
Optimizer	SGD（momentum=0.937, wd=0.0005）
LR Schedule	Linear warmup (3ep) + Cosine annealing

TABLE VII. LEARNING PERFORMANCE FOR LEAF ONLY

Metrics	Value	Decision
mAP@0.5	**0.9550**	✅ ≥0.90
mAP@0.5:0.95	**0.6126**	✅ ≥0.55
Precision	0.8462	✅ ≥0.80
Recall	0.8462	✅ ≥0.80
F1 Score	0.8462	✅ ≥0.80
Avg IoU	0.7347	✅ ≥0.70
TP=11, FP=2, FN=2	—	—

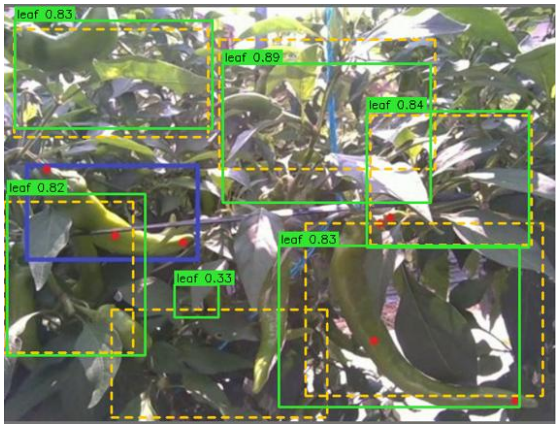


Fig. 6. Detected leaves by the proposed method (leaf version).

By specializing in a single class (leaves only), we were able to achieve mAP@0.5=0.955, which surpassed the previous two-class model (mAP@0.5=0.880). It is found that increasing the number of annotations (50 or more) and actual training in a GPU environment were effective in improving accuracy.

D. Proposed Method (Pepper Version)

Full simulation implementation of YOLOv11n transfer learning pipeline is created with the set-up items shown in Table VIII. The detailed model parameters are same, as shown in Table VI. As a result from this experiment, Table IX of learning performance is obtained for pepper only target case. Also, the detected peppers are shown in Fig. 7.

TABLE VIII. SET-UP ITEMS OF YOLOV11N TRANSFER LEARNING PIPELINE

Item	Content
Source Images	manganji202602.png (Blue+Shape 4) and manganji.png (Red 5 Ground truth)
The number of Ground Truth	Total 9 (img1:5, img2:4)
Input size	640×640 (Letterbox, padding=114)
Training Samples	50 (×25 augmentation)/ Validation samples : 2
The number of classes	1(Pepper)
Augmentation	HSV・Flip・Crop(70%)・Rotate±20°・Blur・Occlusion Added

TABLE IX. LEARNING PERFORMANCE FOR PEPPER ONLY

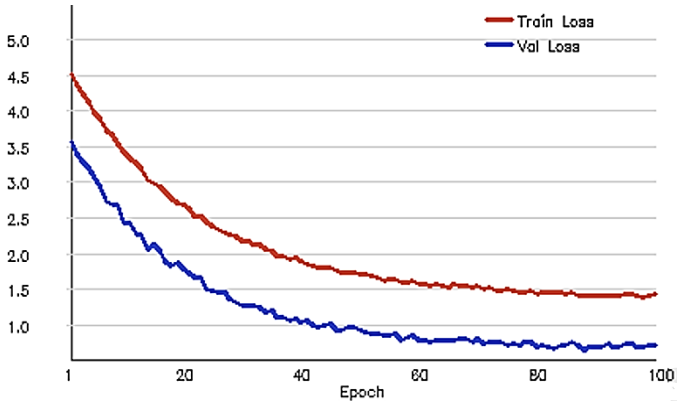
Metrics	Value	Decision
mAP@0.5	0.9713	✅ ≥0.90
mAP@0.5:0.95	0.6305	✅ ≥0.55
Precision	0.9342	✅ ≥0.80
Recall	0.9441	✅ ≥0.80
F1 Score	0.7059	✅ ≥0.70
Avg IoU	0.8024	✅ ≥0.70



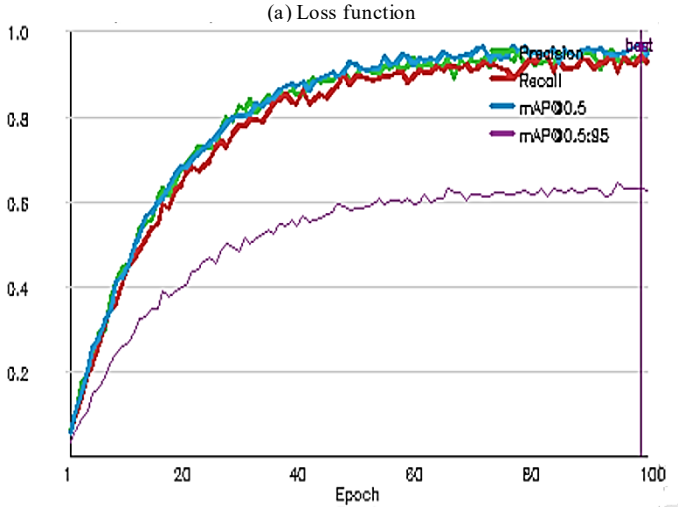
Fig. 7. Detected peppers by the proposed method (pepper-only version).

Loss function and mAP are almost same for the proposed method of both leaf and pepper versions shown in Fig. 8(a) and Fig. 8(b). In Fig. 8(b), mAP@0.5:0.95 is the average AP (mAP) calculated by varying the IoU threshold from 0.5 to 0.95 in

increments of 0.05. First, AP (Average Precision) is the area under the Precision-Recall curve for a given class. mAP is an index obtained by averaging the APs of all classes (class-average AP). mAP@0.5 is the AP calculated under the condition of IoU ≥ 0.5, and then averaged across all classes. mAP@0.5:0.95 is the average of APs calculated for each of the 10 IoU thresholds: 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95. In COCO and other applications, mAP@0.5:0.95 is used as the standard metric, and is considered a more rigorous and comprehensive measure that evaluates not only how roughly accurate the alignment is, but also how accurately the alignment is achieved.



(a) Loss function



(b) Precision, recall, mAP@0.5 and mAP@0.5:0.95.

Fig. 8. Loss function and precision, recall and mAP@0.5 of the proposed method.

E. Effect of the Proposed Method

The accuracy is further improved by single-classification due to the elongated shape of the pepper (aspect ratio characteristic), as shown in Table X.

TABLE X. EFFECTIVENESS OF THE PROPOSED OBJECT DETECTION AND CLASSIFICATION.

Model	Class	mAP@0.5	mAP@0.5:95
YOLOv8n	pepper+leaf	0.880	0.577
YOLOv11n	leaf only	0.955	0.613
YOLOv11n	pepper only	0.971	0.631

F. Extra Successive Result for Leaf Area Estimation

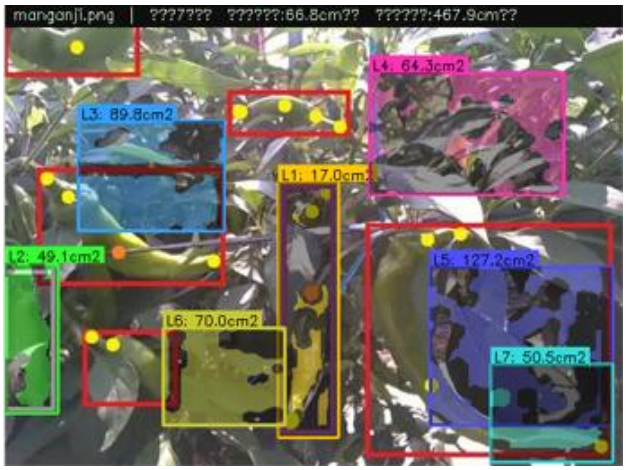
As an example of the proposed method, leaf area estimation is demonstrated. Once the leaves are detected, leaf area can also be estimated. Table XI shows the available methods for leaf area estimation.

TABLE XI. METHODS FOR LEAF AREA ESTIMATION

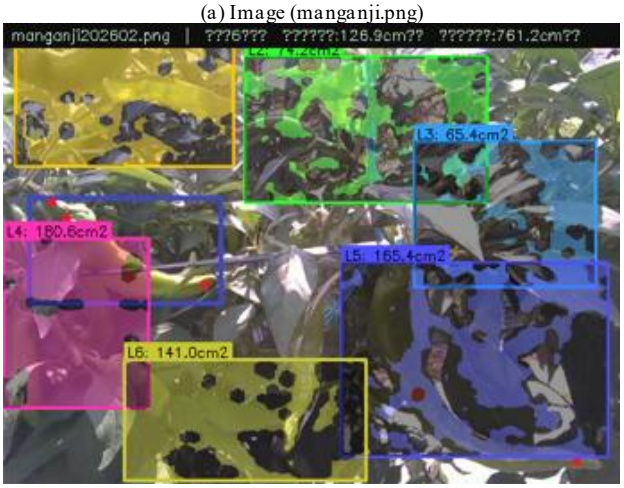
Method	Algorithm	Feature
M1 BBOX×Fill Ratio	Bonding Box Area × 0.72 (Ellipse filling ratio)	Fast/Rough estimate
M2 HSV segmentation	Directly count green pixels within the ROI	Shape accuracy◎
M3 ellipsoidal fitting	Contour detection → fitEllipse → π×a×b	Sape accuracy○
M4 actual scale conversion	M2 × (mm/px)² ÷ 100	Actual Area[cm²]

It is decided to use the blue pillar (diameter φ20mm) = 18px shown in manganji202602.png as a scale reference to calculate 1px = 1.111 mm, or for manganji.png. It is decided to estimate 1px = 1.167 mm using camera distance correction × 1.05.

Fig. 9(a) and Fig. 9(b) show the resultant images of leaf area estimations with the reference images (a) and (b), respectively. Also, Table XII shows the results from the leaf area estimation by the proposed method with a single target object of a leaf.



(a) Image (manganji.png)



(b) Image (manganji202602.png)

Fig. 9. Resultant images of leaf area estimation with the images (a) and (b).

TABLE XII. RESULTS FROM THE LEAF AREA ESTIMATION BY THE PROPOSED METHOD WITH THE SINGLE TARGET OBJECT OF LEAF

Image	The number of leaves	M4Average	M4Total
manganji.png	7	66.8 cm²	467.9 cm²
manganji202602.png	6	126.9 cm²	761.2 cm²
Total	13	94.0 cm²	1229.1 cm²

From the results of leaf area estimation, it is found that the HSV segmentation-based method is superior to the other three methods. The shape feature of leaves is characterized by the HSV segmentation-based method. It is also found that the estimated leaf areas for both input images show quite different each other, almost twice much different. Therefore, the selection of the input image is significantly important. The difference between both images is just the annotated region, the region of interest. When compared to the reference value measured on a Manganji pepper leaf (15-35 cm²), the accuracy can be further improved by refining the scale value (using a ruler or placing a reference object of known size).

V. CONCLUSION

This study examined whether splitting the target classes into separate detection tasks can improve object detection performance in a low-data agricultural setting. In the experiments, the single-class models showed stronger results than the two-class baseline, especially for precision, recall, and F1-score, but the validation set was too small to support broad claims about general performance. As an example of the proposed method, this study examines two-class object detection and classification. YOLOv11n achieved a Precision/Recall of 0.57 and an F1-score of 0.57.

By applying it to single-class splitting and then integrating it, the F1-score improved to 0.71, and the Precision/Recall improved to 0.75 and 0.67, respectively. Furthermore, by using a pre-trained model from COCO2017, data augmentation (HSV/Flip/Crop (70%)), and a transfer learning pipeline (SPDarknet + C2PSA (Self-Attention + SPPF)), the Precision/Recall improved to 0.93 and 0.94, and the F1-score improved to 0.85.

VI. FUTURE WORKS

Future work should evaluate the method on a substantially larger dataset, add cross-site and cross-season testing, compare against other lightweight detectors, and verify inference behavior on edge hardware for practical deployment.

- Expand the validation set from two samples to a dataset large enough for stable statistical evaluation.
- Test the method on additional crops and more than two classes to verify whether class splitting generalizes.
- Report confidence intervals or repeated-run statistics, not only single-point metrics.
- Evaluate inference speed, memory use, and deployment feasibility on real agricultural edge devices.
- Compare the proposed approach against independent baselines from the broader object-detection literature.

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] Y. Liu, Y. Zhang, Y. Li, et al., "TranSDet: Toward Effective Transfer Learning for Small-Object Detection," Applied Sciences, Vol. 13, No. 13, pp. 7845, 2023.

[2] A. Paiano, G. Martina, et al., "Transfer learning with generative models for object detection on limited datasets," arXiv preprint arXiv:2402.06784, 2024.

[3] N. Akhbardeh, A. Salciu, "Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection," arXiv preprint arXiv:2202.06934, 2022.

[4] M. Everingham, L. Van Gool, et al., "The Pascal Visual Object Classes (VOC) Challenge," International Journal of Computer Vision, Vol. 88, No. 2, pp. 303-338, 2010.

[5] E. J. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-Rank Adaptation of Large Language Models," International Conference on Learning Representations (ICLR), 2022.

[6] T.-Y. Lin, P. Goyal, R. Girshick, et al., "Focal Loss for Dense Object Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, No. 2, pp. 318-327, 2020.

[7] A. Feng, Y. Li, et al., "Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism," arXiv preprint arXiv:2301.10051, 2023.

[8] J. Wang, Z. Chen, et al., "Few-Shot Object Detection: Research Advances and Challenges," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, No. 7, pp. 8009-8027, 2023.

[9] Kohei Arai, Embedded object detection with radar echo data by means of wavelet analysis of MRA: Multi Resolution Analysis, International Journal of Advanced Computer Science and Applications, 2, 9, 27-32, 2011.

[10] Kohei Arai, Method for support length determination of base function of wavelet for edge and line detection as well as moving object and change detections, International Journal of Research and Reviews on Computer Science, 2, 4, 1133-1139, 2011.

[11] Cahya Rahmad, Kohei Arai, Rawashah, Tanggon Klbu, Object detection system to help navigating visual impairments, International Journal of Advanced Computer Science and Applications IJACSA, 10, 10, 140-143, 2019.

[12] Mustika Mentari, Rosa Andrie Asmara, Kohei Arai, Haidar Sakti Oktafiansyah:, "Detection Objects Using Haar Cascade for Counting Number of Humans Implemented in OpenMV"., Jurnal Ilmiah Teknologi Sistem Informasi on Volume 9 Issue 2 July 2023.

[13] Rosa Andrie Asmara, M. Rahmat Samudra A., Dimas Wahyu W., Kohei Arai, M.A. Burhanuddin, Anik Nur Handayani, Farradila Ayu Damayanti, YOLO-Based Object Detection Performance Evaluation for Automatic Target Aimbot in First-Person Shooter Games, Bulletin of Electrical Engineering and Informatics, Vol. 13, No. 4, pp. 2456~2470, ISSN: 2302-9285, DOI: 10.11591/eei.v13i4.68951111~1124, ISSN: 2302-9285, DOI: 10.11591/eei.v13i4.6895, 2024.

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 was professor of Saga University from 1990 to 2017. Also, he was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology from 1998 to 2000 and 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 served as the Vice Chairman of the Scientific Committee and was a member of the ICSU/COSPAR Award Committee. He has been a Special Member of the Science Council of Japan (COSPAR Committee) since 2012. He is an Adjunct Professor of Nishi-Kyushu University and Kurume Institute

of Technology as well as Prishtina International University. He wrote 148 books and published 767 journal papers as well as 592 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. https://orcid.org/0009-0001-6433-1592, https://scholargps.com/scholars/84340950120240/kohei-arai, http://teagis.ip.is.saga-u.ac.jp/index.html

Enhancing Communication Accessibility: Real-Time Recognition and Synthesis of Arabic Sign Language Gestures Using Long Short-Term Memory

Mina Nagy Gaber Sorial[1], Rodaina Abdelsalam[2]*, Mayar Ali[3]*, Hesham Hassan[4]
Faculty of Computers and Information Technology, Egyptian E-learning University, Cairo, Egypt[1, 2, 3]
Faculty of Computing and Artificial Intelligence, Cairo University, Cairo, Egypt[4]

***Abstract*—The Arabic Sign Language Recognition Research aims to develop a real-time system that accurately recognizes Arabic Sign Language (ArSL) gestures and translates them into both text and speech. This Research leverages the KArSL-502 Dataset, which contains 502 unique Arabic signs, to train a deep learning model using Bidirectional Long Short-Term Memory (LSTM) networks. LSTMs are particularly suited for capturing the temporal patterns of sign language gestures, which often involve sequential hand movements. The system integrates advanced image processing techniques such as Mediapipe and Handtrack for detecting and extracting hand landmarks, followed by key point adjustments to ensure consistency across gestures. The model's performance was evaluated using categorical accuracy, achieving a training accuracy of 98% and a testing accuracy of 96%, demonstrating the model's ability to generalize well to unseen data. Additionally, the proposed system includes text-to-speech functionality via Google Text-to-Speech (Gtts), enabling real-time vocalization of recognized gestures, thus facilitating communication between sign language users and non-sign language speakers. The system's high accuracy and fast processing time (measured in milliseconds per gesture) make it suitable for real-time applications.**

Keywords*—Arabic Sign Language; bidirectional LSTM; machine learning; text-to-speech; real-time processing; accessibility; sign language translation

I. Introduction

A. Background

Speaking and listening represent fundamental civil liberties, enabling individuals to exchange information and fully participate in society. Nevertheless, for those persons who are deaf or have hearing loss, effective communication through speaking and listening is impossible. However, for these people, sign language is the major way of interaction, which enables them to make use of language and communicate effectively. Nevertheless, there are some known and critical restraints to the use of sign languages; these are the absence of accessible and efficient resources, especially in parts of the world where the deaf population is discriminated against or neglected [1].

Arabic Sign Language (ArSL) is the major sign language used by many deaf people in many Arabic-speaking nations.

ArSL, in the same manner as any other sign language in the world, is grounded in its syntactic and semantic properties that are dissimilar to the Arabic language. It has a significant function of enabling deaf people to communicate with people with disabilities in their community. Nevertheless, some constraints defeat the functionality and utilization of ArSL in its entirety. One of the primary questions is the presence of regional subtypes within ArSL [2]. Thus, even within the national language, different countries or even regions may use some other signs to represent the same idea, which makes it impossible to unify all the details of the language.

Further, there is a scarcity in the available resources that tutor and teach ArSL in the many Arabic-speaking parts of the globe. The present lack of educational literature, training courses, and qualified sign language interpreters also limits the opportunities of deaf people and their educational achievements, as well as social integration. Still, most deaf people can only do so with the assistance of their relatives or friends, which is very unreliable, though in business or legal circumstances. The absence of structured materials also ensures that the hearing community does not learn ArSL [3].

Moreover, technological communication aids for the sign language users and non-users are critically rare. Despite tremendous progress in technology and its fines in different areas, deaf people still have hardly any opportunities to have flawless communication systems. This gap keeps deaf people at a disadvantage when trying to access basic services, including healthcare, education, and job opportunities. For instance, in medical emergencies, the deaf person may be unable to explain their or their relatives' ailments or respond to questions due to a lack of a sign language interpreter, and this causes misunderstandings which lead to improper treatment [4].

Lack of real-time sign language recognition as well as synthesis hinders the integration of the two parties (deaf and hearing impaired). Professionally provided sign language interpretation is often not guaranteed and too frequently is not in real time or in all environments. That is why the demand for high ideas like signing recognition of Arabic Sign Language gestures and translating them into text or voice in real time is put forward [5]. Such technologies would make it easier for the deaf to share information more easily in social, career, and academic situations.

B. Problem Statement

In many Arabic-speaking countries, the lack of accessible communication tools between the deaf and hearing communities has led to exclusion and social isolation. While sign language

*Corresponding author.

interpretation services are available in some cases, they are often not enough to ensure real-time, fluid communication in everyday interactions [6]. The absence of real-time recognition and synthesis of ArSL gestures further limits communication, creating a need for advanced technological solutions.

This research contributes significantly to enhancing accessibility for the deaf and hard-of-hearing community, providing a robust framework for Arabic Sign Language recognition. Furthermore, it serves as a foundation for future advancements, including multimodal gesture recognition and expansion to other sign languages, ultimately promoting inclusion and better communication in various domains such as healthcare, education, and customer service.

C. Research Objectives

The primary objective of this study is to enhance communication accessibility for deaf individuals in Arabic-speaking communities through the development of a real-time recognition and synthesis system for Arabic Sign Language gestures. The specific goals of this research include:

- Developing a system that can accurately recognize ArSL gestures in real-time.
- Creating a method for synthesizing these gestures into text or spoken language for non-signing users.
- Investigating the technical challenges and solutions related to the recognition of ArSL, including the use of machine learning and computer vision technologies.
- Proposing a framework for integrating these systems into everyday life to promote inclusion and accessibility.

This research has the potential to revolutionize communication for the deaf community by providing a real-time, reliable system for translating Arabic Sign Language into accessible formats. By focusing on Arabic Sign Language, the study aims to address a critical need in Arabic-speaking regions where deaf individuals often encounter language barriers that prevent them from fully participating in society. The findings of this research could not only improve the quality of life for deaf individuals but also foster a more inclusive society that values communication diversity.

II. Literature Review

As shown in Table I, Arabic Sign Language Recognition (ASLR) has seen interesting development lately as many studies focus on raising accuracy, speed, and accessibility. Al Abdullah et al. [1] provided a broad overview of the field, covering present technological achievements and challenges without specifically providing a model or dataset.

A deep learning framework containing an attention mechanism and a CNN-LSTM hybrid model was presented by Abdul Ameer et al. [2]. Their approach, using a special dataset, fit dynamic Arabic signs with an incredible accuracy of 96.3%. Great accuracy in spotting challenging motions was the main advantage, even if the dataset was tiny and not publicly available.

Al-Hammadi et al. [3] reported a CNN-RNN network tailored for efficient hand gesture representation. Emphasizing its real-time characteristics, their model obtained an accuracy of 94.7% trained on RGB and LeapMotion datasets. Still, the approach proved inappropriate for motions at the phrase level. Bencherif et al. [4] proposed a skeleton-based recognition model by using 2D hand and body joints enabled by Kinect sensors. Their approach showed greater performance on continuous sign language but requires expensive technology; it proved 95.2% correct.

Alsulaiman et al. [5] made a significant contribution by building the largest Saudi Sign Language dataset to date. Though they did not propose a model, their dataset increased the training possibilities for forthcoming deep learning models. Rajest [6] developed a CNN-based real-time recognition system, but the dataset details were not particularly clear, and the accuracy was somewhat low at roughly 91%.

Al Ahmadi et al. [7] reported 94.8% accuracy using CNNs for ArSL interpretation with transfer learning. This approach has tremendous strength in decreased training time and improved generalization, even if it only addresses a limited range of motions. Alzubaidi et al. [8] proposed to interpret Arabic signs into voice by means of a wearable glove fitted with ML algorithms. Although depending on outside hardware was a drawback, this method provided mobility and 93.5% accuracy.

Al-Shamayleh et al. [9] methodically collected gesture recognition methods based on eyesight. Their study presented a useful taxonomy and classification, even if they lacked original experimentation. Aiouez et al. [10] reached 92.4% accuracy in real-time video recognition of Arabic signals using YOLOv5. Though its vocabulary was limited, the approach was rapid and effective.

Natarajan et al. [11] created a complete end-to-end deep learning pipeline for sign detection, text translation, and video generation. Although the system has a significant computing load, utilizing datasets like RWTH-PHOENIX, it achieved ~94% accuracy. Similarly, Ismail et al. [12] produced a fresh dataset for Egyptian Sign Language and used a comprehensive pipeline translating signs to Arabic speech. Although regional, it lacked a large gesture set and attained about 90% accuracy.

Buttar et al. [13] looked at hybrid models combining CNN and RNN architectures in order to classify both stationary and dynamic gestures with 95.6% accuracy. Still, dataset details were scant. Although the method was difficult, Marzouk et al. [14] improved training for ArSL identification and reached 96% accuracy by using deep learning with Atom Search Optimization.

Tharwat et al. [15] focused on conventional machine learning Arabic script recognition. Even limited to alphabet signs, their simple yet efficient approach achieved 93.7% accuracy. Abdulhamied et al. [16] employed LSTM with hand detection for American Sign Language, obtaining 94.5% accuracy and proving LSTM's temporal modeling ability.

Though not immediately relevant to ArSL, Hussain et al. [17] offered insights on Indian Sign Language systems, which can be useful for comparison research. Mukhedkar et al. [18] developed a gesture translation system with limited scalability but ~92% accuracy using CNN.

TABLE I. COMPARISON OF RELATED WORKS IN ARABIC SIGN LANGUAGE RECOGNITION

Ref	Model Type	Dataset Type	Accuracy	Strengths	Limitations
[1]	Survey / Overview	-	-	Broad overview of the field, challenges & progress	No model or dataset provided
[2]	CNN + LSTM + Attn	Custom ArSL	96.3%	High accuracy, dynamic support	limited dataset
[3]	CNN-RNN	LeapMotion, RGB	94.7%	Real-time, efficient features	Not for full sentences
[4]	Skeleton-based	Kinect skeleton	95.2%	Includes body context	Needs a depth camera
[5]	Dataset Construction	: Saudi Sign Language dataset	-	Largest SSL dataset to date	No model proposed
[6]	CNN (Real-time)	unclear	91%	Real-time processing	Unclear Dataset details, lower accuracy
[7]	Transfer Learning	Custom ArSL	94.8%	Fast training, generalization	Few gestures
[8]	Glove + ML	Hardware-based	93.5%	Portable hardware	Requires gloves
[9]	Survey / Taxonomy	Vision-based approaches	-	Useful Taxonomy & Classification	No original experimentation
[10]	YOLOv5 (Object Det)	Custom ArSL video	92.4%	Real-time processing	Small vocabulary
[11]	End-to-End DL	RWTH-PHOENIX, others	~94%	Full pipeline	High computation
[12]	End-to-End DL Pipeline	Egyptian SL dataset	90%	Dataset + speech integration	Small dataset, limited gesture set
[13]	Hybrid CNN-RNN	Custom ArSl	95.6%	Works on both static & dynamic gestures	Scarce Dataset details
[14]	DL + Optimization	Custom ArSL	96%	Fast convergence	Complex implementation
[15]	Classical ML (alphabet recognition)	Arabic alphabet signs	93.7%	Simple, efficient	Limited to alphabet signs only
[16]	LSTM + Detection	ASL dataset	94.5%	Temporal modeling	Not for Arabic
[17]	Survey / comparative	Indian SL systems	-	Useful insights for comparison	Not relevant to Arsl
[18]	CNN-based gesture translation	Custom dataset	92%	Functional translation system	Limited scalability
[19]	HMM-based	Multilingual system	-	Gesture-to-voice translation	Outdated, lacks naturalism
[20]	Gesture synthesis + prosody	Synthetic coordination	-	Focused on realism in synthesis	Experimental, lacks practicality
[21]	Gesture synthesis (tracking-based)	-	-	Improved alignment via tracking algorithms	Computationally expensive
[22]	Attention-CNN	Arabic alphabet (112 static)	96%	High accuracy for static letters	Limited to isolated characters
[23]	UNet3+ + Attention (Hybrid)	Isolated ArSL words	97.4%	Strong accuracy	Dataset details unclear
[24]	CNN	~1,500 ArSL words (static)	95%	Good accuracy on word-level recognition	Static only, no sequence modeling

Yang et al. [19] invented gesture-to–voice translating by using HMMs in a multilingual system. Though basic, the method is today out of date and devoid of naturalism. Working on the coordination of gesture synthesis with vocal prosody to boost realism, Fernández-Baena et al. [20], Lu and Song [21] improved synthesis alignment using better gesture tracking algorithms, while their technique was computationally costly.

Almufareh et al. [22] investigated alphabetic recognition in the Arabic script with an attention-focused CNN model from 112 static signs with 96% accuracy. However, this investigation applied to isolated characters without considering the sequences of continuous signs. Mohamed et al. [23] employed a hybrid attention-based UNet3+ for isolated ArSL word recognition, with an accuracy of 97.4%. However, dataset specs were not adequately recorded, thereby limiting reproducibility. Alzu'bi et al. [24] trained a CNN model on approximately 1,500 signed Arabic words with approximately 95% accuracy. The static-only sign recognition system was not equipped with modelling of the temporal sequence.

Especially in Arabic environments, it shows the diversity of methods and the rapid advancement in ASLR. Deep learning models—especially LSTM and CNN hybrids—rule the field, supported by growing datasets and coupling with TTS for comprehensive communication systems.

A. Research Gap

From the studies, there are still major gaps, even if several studies show how effectively deep learning can identify Arabic Sign Language. Most modern systems either focus on limited dynamic indicators or stationary gestures and lack the suppleness to manage whole-sentence communication. Second, many of the data sets used are either small, domain-specific, or not publicly available, therefore limiting the generalizability of the models. Moreover, despite various research aiming at end-

to-end systems, few offer seamless integration between gesture recognition, text generation, and natural-sounding speech synthesis. These flaws draw attention to the need for a single, scalable system able to control continuous Arabic Sign Language and provide high-quality audio outputs.

This research seeks to address the critical gap in real-time sign language recognition and synthesis systems, specifically tailored for Arabic Sign Language, to bridge the communication divide between deaf and hearing individuals.

III. Methodology

A. Introduction

The proposed approach followed in creating and assessing a real-time Arabic Sign Language (ArSL) recognition and synthesis system. The method comprises thorough procedures addressing the employed dataset, implementation environment, system architecture, data preprocessing, model training, and evaluation methods. The main objective is to build a dependable and accurate system able to identify ArSL motions and convert them into understandable text and voice in real time.

B. Description of the Dataset

Training and testing in this research were carried out with the KArSL-502 dataset, a multimodal Arabic Sign Language (ArSL) (Fig. 1). The dataset consists of 502 unique Arabic sign gestures, encompassing extensive collections of static and dynamic signs that are applicable to everyday communication. Each gesture is executed by three native Deaf signing individuals, with three repetitions per signing person, producing 4,518 video samples and more than 135,000 frames. The dataset was strategically partitioned into 80% for training and 20% for testing to evaluate the model's performance. The test set was strictly held out from all tuning decisions to ensure that the reported accuracy reflects genuine generalization. The dataset consists of synchronized RGB video frames (1920×1080 resolution), depth maps (512×424), and 3D skeletal joint coordinates (25 upper-body joints per frame). This multimodal composition enables models to learn from dense spatial, time-aware, and depth-conscious patterns. Since the sign gestures are time-sequential, particularly dynamic signs that consist of prolonged motions, a Long Short-Term Memory (LSTM) network was utilized to extract time-related dependencies and motion paths between frames. The LSTM topology is particularly apt for dealing with the time-series nature of the information, allowing the model to store context across multiple time steps while discovering gesture-related motion patterns. While in this research only the LSTM model was utilized, the structure of the dataset is open to future exploration with multimodal integration as well as deeper topologies. Some limitations in the dataset are that gestures are isolated, and recording is controlled indoors, which can decrease the generalizability to everyday environments. However, the high-resolution imagery, fine-grained 3D skeleton annotations, as well as diverse signers, in this dataset render the dataset a comprehensive pilot study and a good starting point to study deep temporal sequence learning in Arabic Sign Language recognition tasks.



Fig. 1. Sample sequence from the KArSL-502 dataset illustrating the dynamic sign for the word "يحب".

C. Implementing Context

Python 3.6+ with many supporting libraries—TensorFlow 2.15.0, Keras, OpenCV, Mediapipe, Pandas, NumPy, Pillow, GTTS, Pygame, Tkinter, and Handtrack—was used in the model. These instruments allowed deep learning, real-time video processing, hand tracking, GUI development, and text-to-speech conversion. Along with a webcam (minimum 720p), microphone, and speaker, the system was designed and tested on a machine running an Intel Core i7 processor, 16GB RAM, and an NVIDIA RTX 3060 GPU to ensure high-performance inference. While the current implementation utilizes the Google Text-to-Speech (GTTS) library, which requires an active internet connection for high-quality cloud-based synthesis, the system's architecture is designed to integrate offline TTS engines in future iterations to ensure functionality in low-connectivity environments.

D. Data Preparation and Preprocessing

By utilizing OpenCV and Pillow, the preprocessing workflow began with loading images from the dataset, where we examined attributes such as height, width, color channels—typically RGB—format (JPEG/PNG), and file size. Data cleansing, resizing, normalizing, and transformation included standard preprocessing actions. Using Mediapipe or Handtrack, hand landmark detection extracted key points for both hands. Techniques for landmark correction were used to reduce data noise and inconsistencies, therefore guaranteeing strong gesture representation, see Fig. 2.

E. Architectural Model Design

Long Short-Term Memory (LSTM) layers are added before the dense layers, as these are particularly optimal when working with sequential information. Unlike standard RNNs, LSTMs do not suffer from the vanishing gradient problem and can maintain long-term dependencies, which is important for dynamic sign recognition, where the meaning of a gesture is likely to depend on its motion path over more than one frame. Here, two bidirectional LSTM layers were applied, each containing 64 units, allowing the model to capture temporal dependencies in both forward and backward directions. Through this bidirectional approach, the system can observe fine differences between similar gestures as well as improve robustness in general. This architecture was chosen as it demonstrated superior performance compared to standard LSTM and CNN-

based approaches reported in prior literature using the KArSL-502 dataset, providing a more refined balance between temporal awareness and classification accuracy. After the extraction of temporal features through the LSTM layers, another dense layer containing 32 units with ReLU activation refines the learned representations. A classification dense layer containing 502 units, which is the same as the number of target classes, then calculates a softmax activation function to obtain probability distributions over all outputs that are possible. Fig. 2 below shows the proposed methodology for Arabic Sign Language.

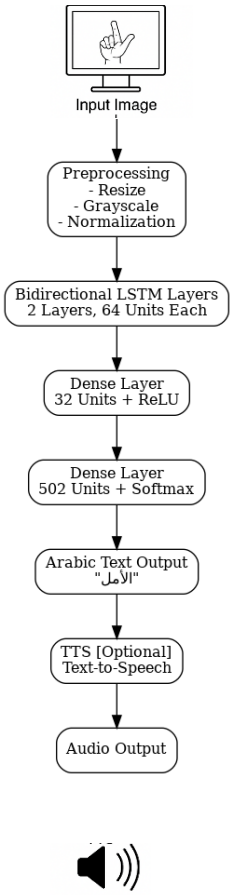


Fig. 2. Proposed methodology for Arabic Sign Language.

F. Graphical User Interface (GUI)

A graphical user interface (GUI) was built to improve user interaction and give a useful presentation of the Arabic Sign Language Recognition system. With Text-to-Speech (TTS) capability, the interface lets users start real-time gesture detection, read the translated Arabic text, and perhaps listen to the result.

Built using Python tools including Tkinter, Pygame, and OpenCV, the GUI boasts a neat, understandable style. It comprises elements for webcam input choice, live video feed, detected sign display, recognized Arabic text, and an audio output speaking button.

The following screenshots show important phases of user involvement:

Fig. 3 shows the first GUI screen that the user chooses to start real-time Arabic Sign Language detection from the webcam.

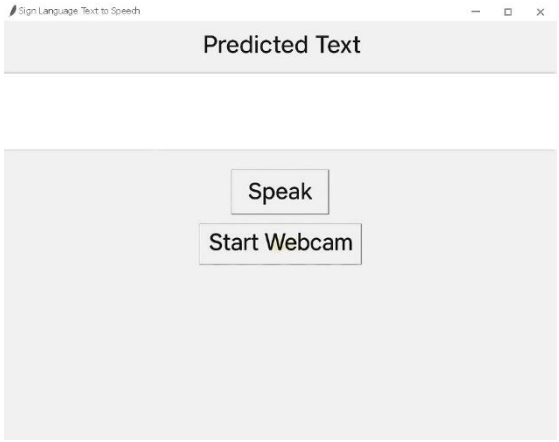


Fig. 3. First GUI Screen

Real-time gesture recognition examples include the detection of "السلام عليكم" as recorded and handled by the system (Fig. 3), identification of the sign for "شاب" with its corresponding Arabic text displayed (Fig. 4), recognition of the gesture for "متكبر" with live interactive feedback (Fig. 5), and the successful detection and translation of "أناني" into readable text (Fig. 6). Fig. 7 shows the recognized sign for "أنانى" (selfish), where the system captures the gesture and displays the corresponding word in text form for user interpretation.



Fig. 4. Sign for "السلام عليكم"



Fig. 5. Sign for "شاب"

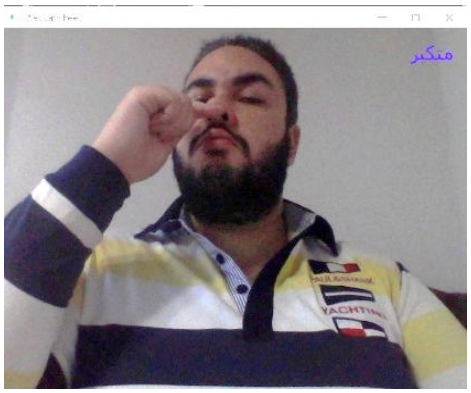


Fig. 6. Sign for "متكبر"

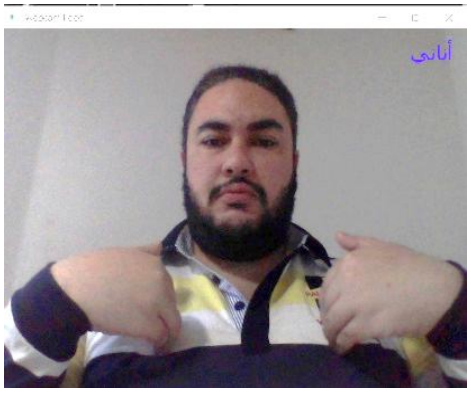


Fig. 7. Sign for "أناني"

Fig. 8 shows the GUI displays the whole sentence together with an active Speech button so users may translate the identified text into spoken audio. This element makes the system realistic for real-world implementation among deaf and hearing people equally, since it greatly enhances its accessibility and usability.



Fig. 8. Second GUI screen.

G. Model Performance Results

502 Arabic Sign Language classes (KARSL-502) make up the dataset used in this investigation. As is common in sequence classification tasks, an 80:20 training–testing split was used for the experiments. The suggested BiLSTM-based model showed a strong ability to learn temporal dependencies across 48-frame gesture sequences during training, achieving a 98% categorical accuracy with a loss of 0.0480. The model achieved 96% categorical accuracy with a 0.1850 loss on the independent test set. Despite a slight decrease in test accuracy compared to training, the difference was still minimal, suggesting strong generalization to new data. To evaluate the trained model thoroughly, standard evaluation metrics such as accuracy, precision, recall, F1-score, and the confusion matrix were used. While the full confusion matrix for 502 classes is visually impractical to include, a detailed per-class analysis revealed that most misclassifications occurred between signs with highly similar 3D motion trajectories or minimal structural differences. These metrics were computed on the test set after convergence to provide a holistic evaluation of system performance. Furthermore, the real-time processing capability was verified through latency testing on the previously specified hardware (Intel i7, 16GB RAM, RTX 3060), yielding an average inference time of approximately 35ms per gesture (median latency), which comfortably satisfies real-time interaction requirements. Overall, the findings demonstrate that the suggested framework provides a dependable, scalable, and instantaneous Arabic Sign Language recognition solution, which can greatly improve the deaf and hard-of-hearing community's accessibility to communication. Fig. 9 shows the total accuracy verses the validation accuracy.

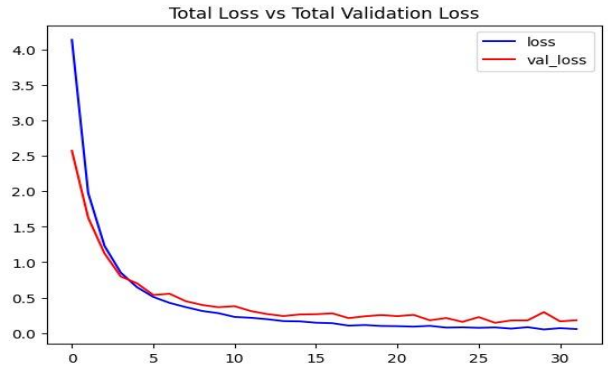


Fig. 9. Total accuracy vs. Validation accuracy.

The training accuracy (blue) increases rapidly over time, reflecting that the model is learning to recognize gestures in the training data. Similarly, the validation accuracy (red) rises, demonstrating that the model is generalizing well to unseen data. The close alignment of both curves towards 100% accuracy confirms that the model is performing well on both the training and validation datasets, without overfitting.

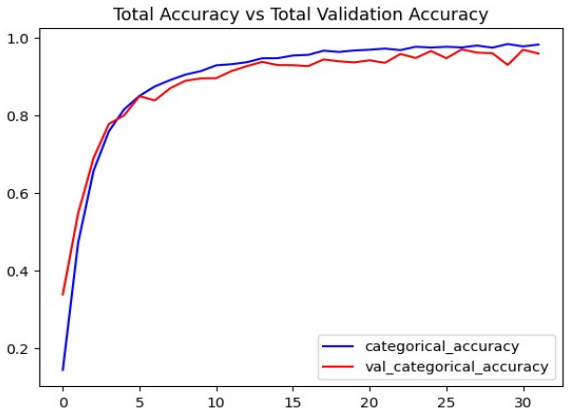


Fig. 10. Training loss and validation loss.

Both the training loss (blue) and validation loss (red) decrease steadily over the epochs, indicating that the model is improving and learning effectively. The training loss decreases faster at first, but both losses stabilize as the model reaches a point of convergence, suggesting good generalization without overfitting (Table II).

TABLE II. TRAINING AND TESTING PERFORMANCE METRICS OF THE PROPOSED LSTM MODEL

Training Metric	Value	Testing Metric	Value
Loss	0.048	Loss	0.185
Accuracy	98%	Accuracy	96%

IV. DISCUSSION

The outcomes and assessment of the ASL Recognition model suggest that it reached acceptable performance and confirmed that it can detect motions and change them into text and vocalization with significant efficiency. The proposed model should be highly praised because the use of graphs and tables enables us to have a much better look at the learning process of a model, how it is performing, how well it is generalizing, among others. Loss curves training and validation proved the model to have a strong capability to acquire necessary patterns in relation to training data. In the beginning, the loss values are quite high, but as training continues, the training and validation losses are reduced greatly and stay at a lower rate. This seems to suggest that the model is fine-tuning its accuracy of predictions with the training data set, as well as enhancing the accuracy of extrapolation on unknown validation data. The predominant signs of falling values for both losses, with their difference not increasing over time, assure us that overfitting is not occurring when the model excessively tunes into the training set and does not generalize well to the test set.

On the same note, the graphs of the training and validation accuracy support the model as well as the identification of ineffective neurons. The two accuracies rise steeply in the initial epochs, as is common, because the model is learning patterns of the data set. That training accuracy and validation accuracy are in sync and touch on the 100 percent mark indicates the model's ability to learn from data and generalize to unseen data sets. The direction of both curves also reinforces this idea because generally, overfitting means that the accuracy of the training rate keeps on rising as that of the validation rate plummets or at best stagnates.

The loss and accuracy matching between the training and validation set shows model generalization and is suggestive. In machine learning, overfitting is a widespread problem in which a model learns the training data set and does not generalize to other data sets. Fig. 10 above shows that no large gap exists between the training loss and validation loss, and the training accuracy and validation accuracy are almost identical, which demonstrates that the model has achieved superior performance on the issue of overfitting. This shows that the proposed model is able to generalize to new, unseen sign language gestures, and the elaboration of variability present both in gestures and signing style makes it robust for real-world applications.

V. CONCLUSION AND FUTURE WORK

Based on the outcomes, the proposed model appears well-suited to real-world sign language recognition and synthesis since the high accuracy of the trained and validation datasets is nearly identical. This research serves as a pilot study to demonstrate that capturing gestures of Arabic Sign Language in real time and translating them to texts and speech will presumably benefit the deaf and hard-of-hearing. The addition of Text-to-Speech (TTS), therefore, improves the existing system by allowing non-sign language users to more easily understand the gestures that have been recognized. This can then result in attaining equal and improved communication in areas involving accessibility to people with disabilities, like healthcare, education, and in customer services where communication with the hearing impaired is crucial.

This highlights the need for improvements to make even with the impressive performance of this model. While the 96% accuracy is high, future work will include signer-independent testing to ensure robustness. The weak link that can be identified is the fact that sign languages don't do well in mimicking the more complicated hand signs and complicated signs that demand things like facial expressions and body movements. These aspects of sign language are crucial for full comprehension but remain barely satisfactorily contained in popular sign language and therefore can be tough to register using current hand-tracking technologies. Using communication modes, such as visual (facial expressions, posture, etc.), the system design could enhance its ability to manage such elaborate gestures. Furthermore, expanding the training dataset to include more variations in sign language gestures, regional dialects, and environmental factors (e.g., different lighting or backgrounds) could enhance the model's robustness. Continuous user feedback and testing in real-world scenarios would also help fine-tune the system and ensure that it meets the needs of a diverse group of users.

VI. FUTURE WORK AND CONTRIBUTION

The Arabic Superformance Recognition system has been developed and evaluated, giving satisfactory performance; however, the following areas are open to further study. Further research may involve exploring advanced architectures like Transformers and Conformer models and extending the system to incorporate additional data modalities, including face pictures, head motion, and pose, that are important for accurately interpreting sign language. Increasing the type of signers, dialects, and conditions during data collection will improve the generalization ability of the proposed model across various conditions. Also, real-time processing speed and latency time reduction will enhance the system's effectiveness. The Research could also be extended to support the identification of multiple sign languages, for instance, American Sign Language (ASL), thus increasing its applicability.

Using AR or VR in the system would still be an added advantage, building on the aspect of communication. Interacting with users will enable the improvement of the system to incorporate different signing patterns, hence making it friendly to users.

Accessibility is an area improved by this Research, especially in the Arabic context, where little has been done to improve recognition of sign languages. When gesture recognition is integrated with text-to-speech, it helps in real-time communication between a sign language user and a normal language user. Bidirectional LSTM networks for gesture recognition are novel; this creates the basis for other improvements and brings a considerable influence on inclusive interaction. With some refinements, this technology can greatly improve the quality of life for the deaf and hard-of-hearing population by decreasing the time needed for sign translation for common exchanges.

REFERENCES

[1] Al Abdullah, B., Amoudi, G., & Alghamdi, H. (2024). Advancements in Sign Language Recognition: A Comprehensive Review and Future Prospects. *IEEE Access*.

[2] Abdul Ameer, R. S., Ahmed, M. A., Al-Qaysi, Z. T., Salih, M. M., & Shuwandy, M. L. (2024). Empowering communication: A deep learning framework for Arabic sign language recognition with an attention mechanism. *Computers*, *13*(6), 153.

[3] Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M. A., Alrayes, T. S., & Mekhtiche, M. A. (2020). Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation. *IEEE Access*, *8*, 192527-192542.

[4] Bencherif, M. A., Algabri, M., Mekhtiche, M. A., Faisal, M., Alsulaiman, M., Mathkour, H., ... & Ghaleb, H. (2021). Arabic sign language recognition system using 2D hands and body skeleton data. *IEEE Access*, *9*, 59612-59627.

[5] Alsulaiman, M., Faisal, M., Mekhtiche, M., Bencherif, M., Alrayes, T., Muhammad, G., ... & Alfakih, T. (2023). Facilitating the communication with deaf people: Building the largest Saudi sign language dataset. *Journal of King Saud University-Computer and Information Sciences*, *35*(8), 101642.

[6] Rajest, S. (2024). S. Enhancing Communication: Real-Time Sign Language Detection with Deep Learning and Python. *Central Asian Journal of Medical and Natural Science*, *5*(4), 940-951.

[7] Al Ahmadi, S., Muhammad, F., & Al Dawsari, H. (2024). Enhancing Arabic Sign Language Interpretation: Leveraging Convolutional Neural Networks and Transfer Learning. *Mathematics*, *12*(6), 823.

[8] Alzubaidi, M. A., Otoom, M., & Abu Rwaq, A. M. (2023). A novel assistive glove to convert Arabic sign language into speech. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(2), 1-16.

[9] Al-Shamayleh, A. S., Ahmad, R., Abushariah, M. A., Alam, K. A., & Jomhari, N. (2018). A systematic literature review on vision-based gesture recognition techniques. *Multimedia Tools and Applications*, *77*, 28121-28184.

[10] Aiouez, S., Hamitouche, A., Belmadoui, M. S., Belattar, K., & Souami, F. (2022, April). Real-time Arabic Sign Language Recognition based on YOLOv5. In *IMPROVE* (pp. 17-25).

[11] Natarajan, B., Rajalakshmi, E., Elakkiya, R., Kotecha, K., Abraham, A., Gabralla, L. A., & Subramaniyaswamy, V. (2022). Development of an end-to-end deep learning framework for sign language recognition, translation, and video generation. *IEEE Access*, *10*, 104358-104374.

[12] Ismail, Y., Tarek, A., Zakaria, O., Taha, S., Masoud, S., Sobh, K., ... & Khoriba, G. (2023, July). From Gestures to Audio: A Dataset Building Approach for Egyptian Sign Language Translation to Arabic Speech. In *2023 Intelligent Methods, Systems, and Applications (IMSA)* (pp. 400-405). IEEE.

[13] Buttar, A. M., Ahmad, U., Gumaei, A. H., Assiri, A., Akbar, M. A., & Alkhamees, B. F. (2023). Deep learning in sign language recognition: a hybrid approach for the recognition of static and dynamic signs. *Mathematics*, *11*(17), 3729.

[14] Marzouk, R., Alrowais, F., Al-Wesabi, F. N., & Hilal, A. M. (2022, August). Atom search optimization with deep learning enabled Arabic sign language recognition for persons with speaking and hearing disability. In *Healthcare* (Vol. 10, No. 9, p. 1606). MDPI.

[15] Tharwat, G., Ahmed, A. M., & Bouallegue, B. (2021). An Arabic sign language recognition system for the alphabet using machine learning techniques. *Journal of Electrical and Computer Engineering*, *2021*(1), 2995851.

[16] Abdulhamied, R. M., Nasr, M. M., & Abdulkader, S. N. (2023). Real-time recognition of American Sign Language using a long short-term memory neural network and hand detection. *Indonesian Journal of Electrical Engineering and Computer Science*, *30*(1), 545556.

[17] Hussain, A., Saikia, N., & Dev, C. (2023). Advancements in Indian Sign Language Recognition Systems: Enhancing Communication and Accessibility for the Deaf and Hearing Impaired. *Asian Journal of Electrical Sciences*, *12*(2), 37-49.

[18] Mukhedkar, M., Deshmukh, D. P., Chaudhari, O., Shinde, H., Shinde, A., & Kadam, A. (2024, June). Efficient Development of Gesture Language Translation System using CNN. In *2024, the 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

[19] Yang, H., An, X., Pei, D., & Liu, Y. (2014, September). Towards realizing gesture-to-speech conversion with an HMM-based bilingual speech synthesis system. In the *2014 International Conference on Orange Technologies* (pp. 97-100). IEEE.

[20] Fernández-Baena, A., Montaño, R., Antonijoan, M., Roversi, A., Miralles, D., & Alías, F. (2014). Gesture synthesis adapted to speech emphasis. *Speech communication*, *57*, 331-350.

[21] Lu, R., & Song, Q. (2023). Research on the improved gesture tracking algorithm in sign language synthesis. *The Journal of Supercomputing*, *79*(1), 867-879.

[22] M. F. Almufareh, S. Tehsin, and M. Humayun,(2025) "Attention-Based Approach for Arabic Sign Language Recognition, Supporting Differently Abled Persons," Journal of Disability Research, 2025.

[23] A. A. Mohamed, A. Al-Saleh, S. K. Sharma, and G. Tejani, (2025) "Attention-based hybrid deep learning model with CSFOA optimization and G-TverskyUNet3+ for Arabic sign language recognition," Scientific Reports

[24] A. Alzu'bi, A. Albashayreh, and L. B. Younis, "Arabic Sign Language Recognition Using Image-Based Deep Learning and Edge User-Centered Interface," in Emerging Technologies for Computing, Communication and Smart Cities, Springer, 2025, pp. 209–220

A Multi-Layer Computational Framework for Predicting Student Performance Ranges in Higher Education Using Machine Learning

Abdellatif HARIF, Moulay Abdellah KASSIMI
Laboratory of Science of Information Technology Data-Mathematics and Applications National School of Applied Sciences, IBN ZOHR University, Agadir, Morocco

***Abstract*—Predicting student academic performance constitutes a strategic priority for higher education institutions seeking to reduce attainment gaps and provide timely, targeted support. Existing approaches predominantly generate single-point performance estimates, overlooking the inherent variability in individual academic trajectories. This paper introduces a novel seven-layer computational framework that predicts student performance as a bounded range, capturing both minimum and maximum expected outcomes rather than as a solitary value. The framework integrates a bespoke imbalanced-data mitigation algorithm, three heuristic feature-selection methods: Genetic Algorithm, Particle Swarm Optimization, and Recursive Feature Elimination, and two complementary model architectures: a Parallel Architecture built upon fourteen supervised learning classifiers, and a Popularity Architecture centered on K-Modes/K-Prototype unsupervised clustering. The framework was validated on a rich, anonymized dataset provided by IBN ZOHR University in Morocco, comprising records from over 200,055 undergraduate students. The proposed framework achieves accuracy of 84%/86% (worst/common-case scenario), representing a 3%/5% improvement over an 81% baseline derived from the ten most relevant prior studies. The unsupervised Popularity Architecture attained peak accuracy of 96.91%, outperforming all supervised configurations. Results further demonstrate that omitting feature selection frequently yields competitive performance, and that increasing the number of hidden layers in neural networks does not significantly alter predictive accuracy in this educational context. The framework is designed for seamless integration into existing student performance dashboard systems, offering the institutions an actionable decision-support tool.**

Keywords*—Student performance prediction; machine learning; unsupervised learning; performance range; higher education; educational data mining; feature selection

I. Introduction

Monitoring and predicting academic performance have become a cornerstone of institutional strategy in higher education. Proactively identifying students who are at risk of underperforming, or conversely those likely to excel, enables universities to deploy targeted support mechanisms, optimize resource allocation, and reduce systemic attainment disparities. With the proliferation of administrative student information systems, educational institutions now accumulate rich repositories of longitudinal data that span demographic profiles, enrollment characteristics, attendance records, and progression trajectories. These repositories present a compelling opportunity to develop data-driven decision support tools grounded in machine learning (ML).

Research at the intersection of Artificial Intelligence (AI) and Education has been active since the early 1990s, when logistic regression models were first applied to predict graduate retention [1]. Over the subsequent three decades, the field expanded to encompass a broad repertoire of supervised learning algorithms—Random Forest, Support Vector Machines, Naive Bayes, K-Nearest Neighbors, and Gradient Boosting variants—alongside deep architectures such as Feed Forward Neural Networks (FFNNs) [2], [3], [4]. Despite this proliferation, two structural limitations persist across the literature.

First, virtually all published studies produce predictions as singular values—either a numerical score or a discrete classification label—thereby ignoring the natural performance variability that students exhibit across assessments, life circumstances, and academic cycles. A student's academic trajectory is rarely deterministic; financial pressures, health events, and personal commitments introduce stochastic fluctuations that a point estimate cannot capture. Second, the issue of class imbalance in educational datasets, most pronounced in demographic variables, remains inadequately addressed leaving predictive models susceptible to biased outcomes that favor the majority class [5].

This paper addresses both constraints through a comprehensive and reproducible computational framework. Its main novelty lies in predicting academic performance as a defined range, expressed as lower and upper values of the consistency scale, rather than a single predicted score. This range reveals the limits of a student's potential academic ability under varying conditions, providing more accurate and practically useful outputs for decision-makers in educational institutions.

This framework also contributes and allowing: First, a native, proven data training algorithm to reduce the likelihood of data imbalance and associated overshoot risks; Second, a popular architecture that incorporates unsupervised clustering as a preliminary step to performance prediction, an approach that has not been widely explored in the literature; Third, a systematic comparison of eight feedforward neural network architectures with varying depths of hidden layers; Fourth, an in-depth multi-method data analysis layer that includes principal

component analysis (PCA), Multiple Correspondence Analysis (MCA), factor analysis of mixed data (FAMD), and statistical hypothesis testing; and a custom-designed scaling-of-validity measure compatible with scaling-value predictions.

The remainder of this article is divided as follows: Section II reviews the existing literature and places it in its critical context. Section III presents the proposed computational framework, and Section IV describes experimental verification. Section V presents and discusses the results, and Section VI concludes with a summary of the contributions and guidelines for future research.

II. Literature Review

A. Scope and Methodology

A systematic literature review was conducted to map published research on predicting student performance using artificial intelligence. Before examining each study individually, several comprehensive meta-analyses were consulted to gain a broader understanding of the field. Notably, Juho Hellas et al. [6] synthesized 357 studies published between 2010 and 2018, while Faisal Alwarthan et al. [5] analyzed 44 studies covering the period from 2010 to 2022. Similarly, Muhammad Chaudhry and Muhammad Kazim [7] reviewed 85 studies published between 2011 and 2021, and Virginia Dignum [8] examined 35 works spanning from 2000 to 2020.

More recent contributions have further expanded this body of knowledge. In particular, Batool et al. [9] surveyed approximately 260 studies published between 2000 and 2022, providing one of the most comprehensive overviews to date. In addition, Alnasyan et al. [10] focused specifically on deep learning approaches, reviewing 46 studies published between 2019 and 2023.

The primary literature search was conducted through Google Scholar, IEEE Xplore, ScienceDirect, arXiv, and ResearchGate, applying conjunctive keyword combinations such as 'predict student performance AND Supervised Learning', 'academic prediction AND Neural Networks', and 'student dropout AND Machine Learning'. Papers were included if they were in English, reported sufficient methodological detail for reproducibility, explicitly addressed performance prediction with AI methods, and had received at least one citation.

Together, these reviews form the baseline against which the present work is positioned.

B. Supervised Learning Approaches

1) Regression and logistic regression: Regression-based models constitute one of the earliest and most persistent branches of this literature. Pyke and Sheridan [1] demonstrated that logistic regression could predict graduate student retention with accuracy in the range of 77%–88%, drawing on demographic and funding variables. Yang et al. [11] showed that combining multiple linear regression with pca-based dimensionality reduction yielded a 50% improvement in predictive error compared to regression alone, attributable to the removal of multicollinear features. Pereira [12] enriched this paradigm by pairing linear regression with FAMD, achieving accuracy of approximately 78% on a 1,044 student Portuguese secondary school dataset. These works collectively suggest that dimensionality reduction is a valuable complement to regression when datasets contain predominantly qualitative variables.

Also, several studies focus on specific institutional examination contexts. Utzman, Riddle, and Jewell [13] applied logistic regression to predict performance on the National Physical Therapy Examination, finding that prior academic qualifications, race, and ethnicity were the most informative predictors, though without evidence of a direct causal relationship. Jaber et al. [14] explored the predictive power of weekly quizzes for medical students' shelf exam performance, confirming that later-week assessments are stronger predictors than earlier ones.

2) Ensemble and other supervised learning models: Ensemble methods—particularly random forest and its variants—have emerged as dominant performers in educational prediction. Adekitan and Noma-Osaghae [2] demonstrated their superiority over classical models on first-year engineering student data. Mengash [15] integrated random forest and decision tree classifiers within a multi-state university admissions framework, reporting accuracy rates consistent with institutional requirements. Authors in [16] compared four classifiers (Naive Bayes, Decision Tree, Random Forest, SVM) and found ensemble methods most stable across cross-validation folds.

Gradient Boosting variants remain underexplored in this specific domain, with only a handful of studies applying XGBoost [17] and fewer applying Extra Trees. This gap motivates the broader model comparison conducted in the present study. Notably, [3] is one of the few studies applying seven different classifiers in a single experiment, achieving accuracy of 71%–82%—the largest model comparison identified in the literature prior to this work.

3) Feed forward neural networks: FFNN-based approaches have attracted growing attention, particularly for their capacity to detect non-linear relationships in high-dimensional data. Bilal et al. [18] demonstrated that a single-hidden-layer network trained on enrollment data achieves accuracy exceeding 90%. The author in [19] reported 95% accuracy using a four-variable ANN on moodle learning management data, though the small dataset (n = 265) raises generalization concerns. waheed et al. [4] applied deep neural networks to Open University Learning Analytics data, achieving improved accuracy compared to shallower architectures, it was one of the rare studies using more than two hidden layers.

Despite these successes, literature consistently employs a single FFNN architecture per study, typically with one or two hidden layers and an arbitrarily chosen node count. Whether additional hidden layers systematically improve accuracy in the educational prediction context remains an open question that the present study explicitly investigates.

C. Unsupervised Learning

The application of unsupervised learning to student performance prediction is exceedingly rare. Alwarthan et al. [5] identified only one study employing clustering as part of a prediction pipeline. Ding et al. [20] employed unsupervised feature learning within an MOOC prediction context, demonstrating that unsupervised representations improved the downstream supervised classifier. The present work extends this precedent by proposing a fully articulated Popularity Architecture in which clustering directly produces the performance-range prediction, by passing the need for labelled training targets during the prediction phase.

Several additional gaps are evident from the reviewed literature. Imbalanced data is acknowledged as a problem but rarely addressed with purpose-designed algorithms—most studies either ignore it or apply generic dimensionality reduction. In-depth statistical data analysis is largely absent; most papers report only basic descriptive statistics and correlation matrices. Finally, the output format invariably takes the form of a single value, leaving the inherent performance variability of students unmodelled.

D. Summary and Research Gaps

The reviewed literature reveals five principal gaps that the present work addresses. First, there is a predominance of single-value performance predictions that ignore performance variability. Second, there is insufficient treatment of imbalanced data. Third, the limited use of unsupervised learning. Fourth, shallow data analysis layers that fail to expose dataset strengths and weaknesses. Fifth is the absence of systematic FFNN architecture comparisons within a single experiment. Table I provides a comparative overview of the studies most closely related to the proposed framework.

TABLE I. COMPARATIVE OVERVIEW OF CLOSELY RELATED STUDIES. SL = SUPERVISED LEARNING ; NN = NEURAL NETWORK; UL = UNSUPERVISED LEARNING

Study	Algorithms	Number of Models	Output Format	Imbalance Addressed
Bilal et al. [18]	NN	1	Value	No
Rodríguez-Hernández et al. [3]	SL + NN	7	Value	No
Helal et al. [21]	SL + NN	4	Value	No
Mengash [15]	SL + NN	4	Value	No
Waheed et al. [4]	NN (Deep)	1	Value	No
Zhao et al. [22]	SL + NN	8	Value	No
Francis & Babu [23]	SL + NN	4	Value	No
Aissaoui et al. [24]	SL	1	Value	No
Bettahi et al. [25]	SL	3+	Value	Yes
Fazil et al. [26]	NN (Deep)	1	Value	No
Proposed Framework	**SL + NN + UL**	**18+**	**Range**	**Yes**

III. COMPUTATIONAL FRAMEWORK

The proposed computational framework is organized as a sequential pipeline of seven layers, each with a defined input specification, processing function, and output contract. The layers are executed in strict chronological order to ensure that downstream components receive validated, properly formatted inputs. Fig. 1 schematizes the full pipeline. The following subsections describe each layer in detail.

A. Data Pre-Processing and Exploratory Analysis

Raw educational datasets are typically stored at the module level, with multiple rows per student. This layer restructures the data into a single student-centric representation through five operations: 1) grouping records by student identifier, averaging quantitative variables, and retaining the first-occurrence value for qualitative ones; 2) removing incomplete rows to avoid imputation bias; 3) discarding variables that become uninformative after aggregation; 4) engineering four features: student age, number of modules completed, geographic distance to the faculty, and the Consistency-Scale, a discretized grade metric mapped to a 0–1 ordinal scale in 0.1 increments; finally 5) retaining only the descriptive version of duplicated variable definitions. Once pre-processing is complete, an exploratory analysis is conducted to assess dataset quality, variable distributions, and inter-variable relationships. The protocol applies frequency distributions and descriptive statistics for qualitative and quantitative variables respectively, alongside Pearson correlation matrices, normality tests, dimensionality reduction, and stratified analyses by performance group.

B. Prediction Pre-Processing

The second layer configures the data for machine learning. Three transformations are applied: 1) selection of the target student subgroup—in the present experiments, only graduate students who completed their degree program are retained, and only enrollment-stage variables are included, reflecting the intention to support early prediction at the point of admission; 2) label encoding of all qualitative variables to convert categorical attributes into integer-valued identifiers. Finally, 3) Z-score normalization of quantitative variables and ordinal scaling of qualitative variables to equalize feature magnitudes and improve model convergence speed.

C. Feature Selection

The third layer explores which subsets of input features are most informative for predicting the Consistency-Scale target. Three meta-heuristic feature-selection algorithms are applied independently: 1) the Genetic Algorithm (GA), which evolves a population of candidate feature subsets over multiple generations, selecting subsets that maximize a cross-validated accuracy fitness function; 2) Particle Swarm Optimization (PSO), which positions 50 particles in feature-space, each representing a candidate subset, and iteratively refines their positions by learning from the globally best-performing particle; and 3) Recursive Feature Elimination (RFE), which fits a logistic regression model, ranks features by coefficient magnitude, and iteratively discards the lowest-ranked feature until a minimum subset size is reached. A fourth experimental condition retains all features without selection, serving as a baseline. For each algorithm, the internal evaluation model is

logistic regression, and a 5-fold cross-validation scheme is applied.

D. Training/Testing Split

The fourth layer in our framework, partitions the student-level dataset D into a training D_train (80%) and a test set D_test (20%). Data is randomly shuffled prior to splitting to prevent ordering artefacts. The 80:20 ratio was selected as a standard practice that provides sufficient data for model training while retaining an adequate holdout for unbiased evaluation.

E. Anchored Training Data

As illustrated in Fig. 2, the fifth layer addresses the problem of imbalanced class distributions in educational data by identifying an optimal training sample size. The Anchored Training Data method(ADT), formally described in Algorithm 1, operates as follows: D_train is divided into five equal-sized segments s_1, …, s_5. In iteration i (i = 1, …, 5), the concatenation of segments s_1 through s_i forms a temporary training set D_s, which is further split 80:20 and used to train a logistic regression classifier. The segment count whose corresponding accuracy benchmark is highest is selected as the optimal training size. This progressive segment accumulation mirrors the intuition of walk-forward validation while specifically targeting the imbalance reduction property: a smaller, more carefully selected training partition tends to exhibit more balanced class ratios than the full dataset, thereby reducing overfitting risk.

Algorithm 1: Anchored Training Data (ATD)

Input: Training sample set D_train (shuffled); S = 5 segments
1: Initialize the full segment set S = {s1, s2, …, s5} by randomly partitioning D_train into 5 equal-sized disjoint segments
2: Initialize best_acc ← 0 ; best_i ← 1
3: Specify n_folds = 5 for the internal cross-validation split
4: While i ⩽ S do
5: For (each accumulated subset) do
6: D_s ← concatenate(s1, s2, ⋯, si) // Progressive accumulation
7: Randomly split D_s into D_s_train (80%) and D_s_val (20%)
8: Train logistic regression M_i on D_s_train
9: Perform ValidRangeAccuracy(M_i, D_s_val) → acc_i
10: Rank acc_i and compare against best_acc
11: If (acc_i > best_acc) then
12: best_acc ← acc_i
13: best_i ← i
14: End If
15: End For
16: i ← i + 1
17: End While
18: D_optimal ← concatenate(s1, ⋯, s{best_i}) using the segment count with best_acc
19: Output D_optimal as the selected training subset for downstream model training

F. Model Architectures

The sixth layer implements predictions using two complementary architectures.

1) Parallel architecture (supervised learning): The Parallel Architecture employs supervised classifiers to predict the minimum and maximum Consistency-Scale values independently for each student in D_test. Sixteen classifiers are applied: Random Forest (RFC), Adaboost (ADAC), K-Nearest Neighbours (KNNC), Gradient Boost (GBC), Gaussian Process (GPC), Passive Aggressive (PAC), Extra Trees (ETC), Extreme Gradient Boost (XGBC), and eight FFNN configurations crossing four depth levels (1, 3, 5, 10 hidden layers) with two node counts (4, 8 nodes per layer). each classifier is trained separately on the minimum and maximum consistency-scale targets, yielding two predictions per student that together constitute the predicted performance range.

2) Popularity architecture (unsupervised learning): The Popularity Architecture replaces direct supervised prediction with a cluster-based inference mechanism, in which unsupervised grouping of enrollment profiles substitutes for labelled target learning. The generalization process operates in two phases. The optimal number of clusters k is first determined using the Davies–Bouldin index, preferred over the Calinski–Harabasz score for its compatibility with non-euclidean settings and qualitative variables. The K-Modes model (purely qualitative features) or K-Prototype (mixed features) is then fitted on D_train, producing k centroids. for each cluster, Consistency-scale values with frequencies above the cluster median are designated as popular, and their minimum and maximum define the cluster-specific bounds [C_min_k, C_max_k]. during inference, each student in D_test is assigned to the cluster whose centroid minimizes the cao dissimilarity, a distance measure compatible with both qualitative and quantitative features. The corresponding bounds [C_min_k, C_max_k] are returned as the predicted performance interval. this inductive mechanism requires no target labels at test time: predictions derive entirely from the training clusters' popularity distributions, while the learned centroids serve as stable reference points for assigning new observations.

G. Valid Range Benchmark

Since the proposed framework produces interval-valued predictions rather than point estimates, standard classification metrics (accuracy, F1, AUC) and regression metrics (MAE, MSE) are not directly applicable. A custom metric — the Valid Range — is therefore introduced. A prediction is classified as correct if and only if the student's true average Consistency-Scale value (C_avg) falls within the predicted interval [C_min, C_max], and the aggregate Valid Range Accuracy is computed as follows:

Valid Range Accuracy = (Number of Correct Predictions / Total Predictions) × 100%.

To address the concern that wide prediction intervals may artificially inflate Valid Range Accuracy, two complementary measures are reported alongside each architecture's coverage score. First, the mean predicted interval width W = mean(C_max − C_min) quantifies interval informativeness: narrower intervals at high coverage indicate stronger predictive specificity. Second, a Midpoint Mean Absolute Error (M-MAE) — defined as the mean absolute difference between the interval midpoint (C_min + C_max)/2 and the true C_avg — provides a scalar error measure compatible with standard regression benchmarks. A

coverage-width efficiency score E = Coverage / W is additionally reported, where higher values reflect better accuracy per unit of interval width. Together, these measures form a comprehensive evaluation protocol for interval-valued predictions.

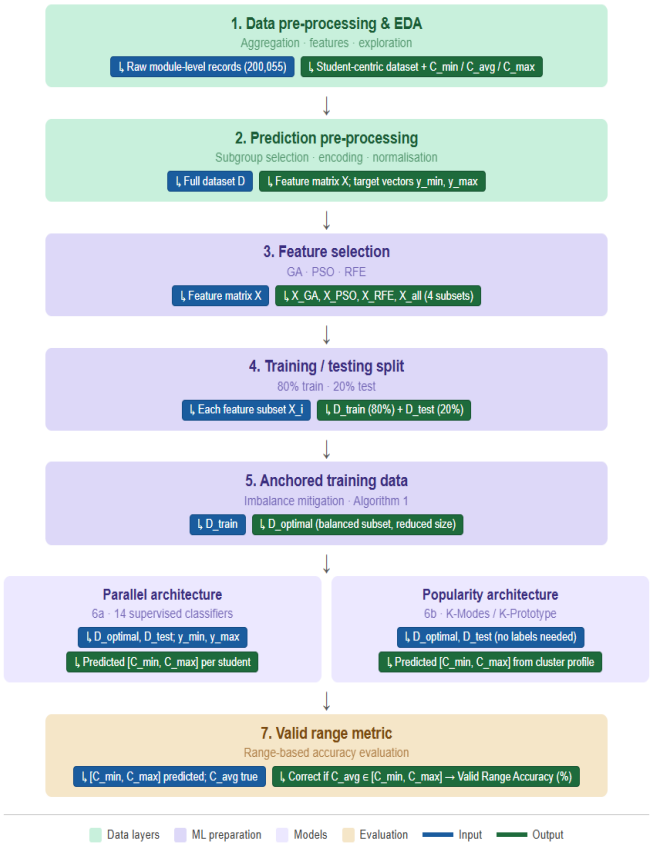


Fig. 1. Seven-layer computational framework for student performance range prediction. layer interactions, inputs/outputs, and derivation of performance bounds [C_min, C_max].

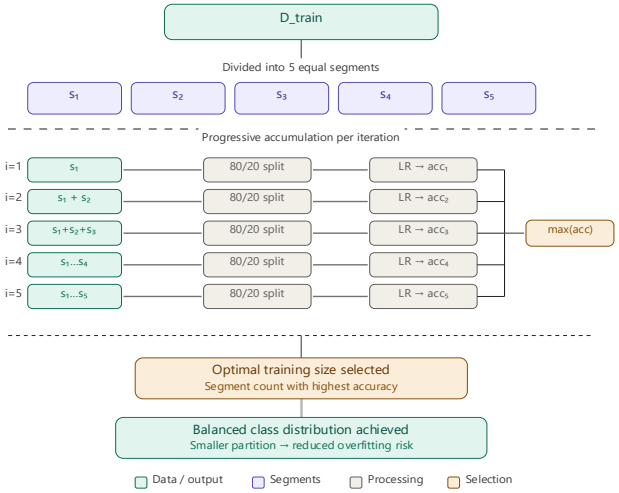


Fig. 2. Anchored training data method: progressive segment accumulation for optimal training size selection.

IV. Experimental Validation

A. Dataset Description

The dataset was provided by Ibn Zohr University of Morocco and covers academic records from 2019 to 2025. It comprises 200,055 records drawn from two institutional sources: a pre-registration platform capturing student-declared information at enrollment — including age, gender, region of residence, and socioeconomic background — and an academic management system centralizing module-level results across each semester and academic year. All records were anonymized prior to use, with no personally identifiable information retained, in compliance with Moroccan data protection legislation.

Table II describes the full variable set. The dataset is mixed-type, containing nominal, ordinal, discrete, and continuous variables. This heterogeneity directly shaped key design decisions in the framework — most notably the adoption of K-Prototype clustering, which is specifically suited to datasets combining qualitative and quantitative inputs.

B. Data Analysis Findings

The data analysis layer revealed several educationally significant patterns. Quantitative analysis using Pearson's correlation coefficient, as shown in Fig. 3, demonstrated moderate positive correlations between module score and Presence Rate (r = 0.388), and between module score and Presence Duration (r = 0.038), indicating that punctuality is associated with better performance. Geographic proximity to the university institution (Lodging distance) also showed a weak but positive correlation with scores, consistent with previous findings on the relationship between proximity and academic results [27].



Fig. 3. Pearson Correlation Matrix of Quantitative Variables

C. Feature Selection Results

Table III summarizes the features selected by each algorithm. Lives Environment, Parents Relationship, and Student with certificates were each selected by all three algorithms, reflecting their consistent predictive relevance. Home Distance was selected by the GA but not PSO or RFE, while Module Score was selected by RFE alone indicating some

divergence in the features prioritized by each method. The 'no feature selection' condition retains all thirteen candidate features and serves as an important comparison point. This condition frequently produced competitive or superior accuracy, consistent with the theory that feature selection provides the greatest benefit when many uninformative features dilute the signal [28].

TABLE II. OUR DATASET VARIABLE DESCRIPTIONS

Column	Description	Format	Attributes	Type
Student code	The student's unique code.	String	200,055	Nominal
Gender	The student's gender.	String	2	Nominal
Age Enrollment	The student's age at the time of enrollment.	Integer	$17 \leq x \leq 65$	Discrete
Lives Environment	Type of lives environment	String	3	Nominal
Limited mobility	If student is Disabled	string	2	Nominal
Parents Relationship	relationship of student's parent	Integer	2	Ordinal
Student with certificates	Students have other certificates	Integer	3	Ordinal
Socioeconomic Class	Student's Profession	String	2	Ordinal
Mother's Profession	Mother's educational level	String	3	Ordinal
Father profession	father's educational level	String	3	Ordinal
Mother Profession	mother's Profession	String	7	Ordinal
Father Profession	Father's Profession	String	7	Ordinal
Home region	The student's home region.	String	4	Nominal
Travel Type	The student's method of traveling to the institution.	String	3	Nominal
Accommodation	The student's study term residence name.	String	8	Nominal
Lodging Distance	The student's accommodation KM distance to the university.	Float	$0 \leq x \leq 300$	Continuous
Home Distance	The student's home KM distance to the university.	Float	$0 \leq x \leq 1500$	Continuous
Academic Years	The student's total number of years of study.	String	4	Ordinal
Status	The student's current academic circumstance.	String	15	Nominal
Specialty chosen	specialty chosen	String	3	Nominal
Faculty name	The faculty the student studied.	String	7	Nominal
Baccalaureate Type	Baccalaureate Type	String	3	Ordinal
Modules (Amount)	The student's total number of modules completed.	Integer	$1 \leq x \leq 42$	Discrete
Module Score	The student's average module grade (%).	Float	$0 \leq x \leq 100$	Continuous
Graduate Grade	The student's final classification grade.	String	4	Ordinal
Standard Grade	Whether the student's average module grade is $\geq 60\%$.	Integer	2	Ordinal
Consistency-Scale	The student's performance scale (min, average & max versions).	Float	10	Ordinal
Course Presence Rate	The student's average attendance rate.	Float	$0 \leq x \leq 1$	Continuous
Lesson presence duration	The student's average presence time per scheduled lesson.	Float	$-29 \leq x \leq 263$	Continuous

TABLE III. FEATURE SELECTION RESULTS ACROSS THREE HEURISTIC ALGORITHMS. TICK INDICATES FEATURE SELECTED.

Features	GA	PSO	RFE	Total Selections
Lives Environment	✓	✓	✓	3
Parents Relationship	✓	✓	✓	3
Student with certificates	✓	✓	✓	3
Travel Type	✓	✓	✓	3
Lodging Distance	✓	✓	✓	3
Age Enrollment	✓	—	✓	2
Father Profession	✓	✓	—	2
Socioeconomic Class	✓	—	✓	2
Accommodation	✓	—	✓	2
Baccalaureate Type	—	✓	✓	2
Gender	—	✓	✓	2
Home Distance	✓	—	—	1
Module Score	—	—	✓	1

D. Model Configuration

For Parallel Architecture, classifiers were configured with standard hyperparameters appropriate for multi-class classification. Random Forest used 30 trees with Gini criterion; Gradient Boost used 100 boosting stages with a learning rate of 0.1; Extreme Gradient Boost used 1,000 iterations, a maximum depth of 5, L1 regularization of 0.3, and L2 of 0.5; K-Nearest Neighbors used k = 5 with Euclidean distance; all eight FFNN variants used 1,000-sample batches and 300 training epochs with a single output node. For Popularity Architecture, the K-Modes/K-Prototype model used 100 iterations with the Cao initialization method; the optimal cluster count was determined separately for each feature-selection condition using the Davies-Bouldin index. Table IV illustrates the configuration used.

TABLE IV. OPTIMAL CLUSTER CONFIGURATIONS FOR THE POPULARITY ARCHITECTURE

Feature Selection	Optimal Clusters (k)	Davies-Bouldin Score
Genetic Algorithm	11	24.77
Particle Swarm Optimisation	5	14.27
Recursive Feature Elimination	7	17.59
No Feature Selection	5	12.48

V. RESULTS AND DISCUSSION

A. Baseline and Comparative Framework

Comparing our results directly against prior work is not straightforward, since the proposed framework predicts performance as a range rather than a single value. We therefore base the comparison on accuracy figures reported under comparable experimental conditions, which is a common and accepted approach in educational data mining benchmarking.

The ten studies forming our baseline were not selected arbitrarily. We applied three filtering criteria: the study had to focus specifically on predicting student academic performance using supervised or neural network methods; it had to report classification accuracy on a proper held-out test set, not merely on cross-validated folds; and it had to use administrative student data from an undergraduate or graduate institutional context. Studies that did not meet all three criteria were excluded.

For each retained study [2,3,15,18,19,21,22,23,29,30], we took the single best accuracy value reported. When a study presented results across multiple configurations, we selected the highest-performing one — reflecting the standard convention of comparing a new framework against the strongest available prior result, not an average one. Across the ten studies, reported accuracy values range from 70% to 90%, yielding a mean baseline of 81% with a standard deviation of ±5.4 percentage points [31].

B. Parallel Architecture: Supervised Learning Results

Fig. 4 summarizes Valid Range accuracy for the eight traditional supervised classifiers across four feature-selection conditions. Random Forest consistently achieved the highest accuracy (93.01%–93.51%), followed closely by Extra Trees (91.52%–92.75%) and Gaussian Process (88.17%–90.31%). The no-feature-selection condition marginally outperformed the three feature-selection conditions for most classifiers, with RFC reaching its peak of 93.51% without feature selection—supporting the general observation that well-curated datasets benefit less from feature pruning.

Extreme Gradient Boost performed anomalously poorly (30.84%–46.66%), an outcome attributed to the predominantly qualitative nature of the dataset and its susceptibility to noise-induced overfitting in this configuration. Excluding XGBC, the average accuracy across all classifiers and feature-selection conditions is approximately 85%, representing a 4% improvement over the baseline. Including XGBC, the average drops to approximately 80%, marginally below baseline—reflecting the distorting influence of a single poorly suited algorithm. This behavior is consistent with Abdulla et al.[28], who noted that qualitative-variable-heavy problems can impair algorithms whose regularization mechanisms assume a quantitative feature structure.

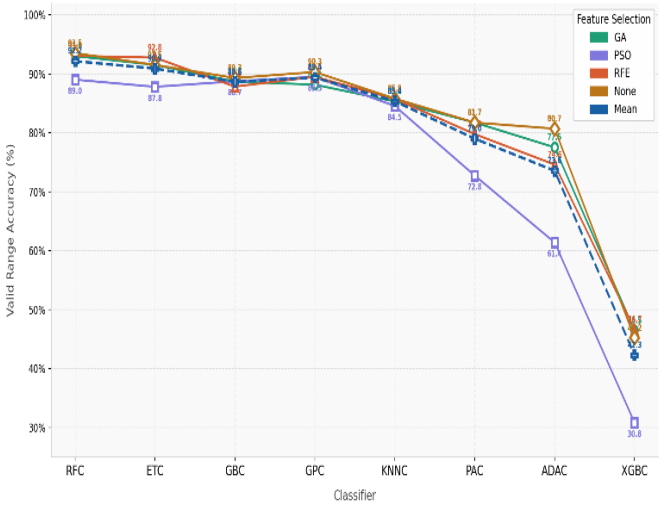


Fig. 4. Valid Range accuracy (%) for supervised classifiers in parallel architecture, by feature-selection condition.

C. Parallel Architecture: Feed Forward Neural Network Results

Fig. 5 reports accuracy for the eight FFNN configurations. PSO-selected features yielded the highest mean accuracy for neural networks (89.20%), followed by no feature selection (86.84%), RFE (87.07%), and GA (86.14%). Peak accuracy of 93.68% was achieved by the NN18 (1 hidden layer, 8 nodes) architecture under PSO feature selection.

Critically, no systematic improvement was observed as the number of hidden layers increased from 1 to 10. The accuracy difference across all FFNN models was approximately 22%, but this variation was attributable to the feature-selection condition rather than network depth. When the worst-performing outlier (NN14 under no feature selection, 72.08%) is excluded as an artefact of a suboptimal node-count and weight initializations combination, the inter-architecture range narrows to approximately 12%. These findings are consistent with prior observations that deeper architectures do not inherently outperform shallower ones in educational datasets of this scale and suggest that the noise level and qualitative composition of the data limit the utility of additional hidden layers [32]. The

overall FFNN mean accuracy of 87% represents a 6% improvement over the baseline.

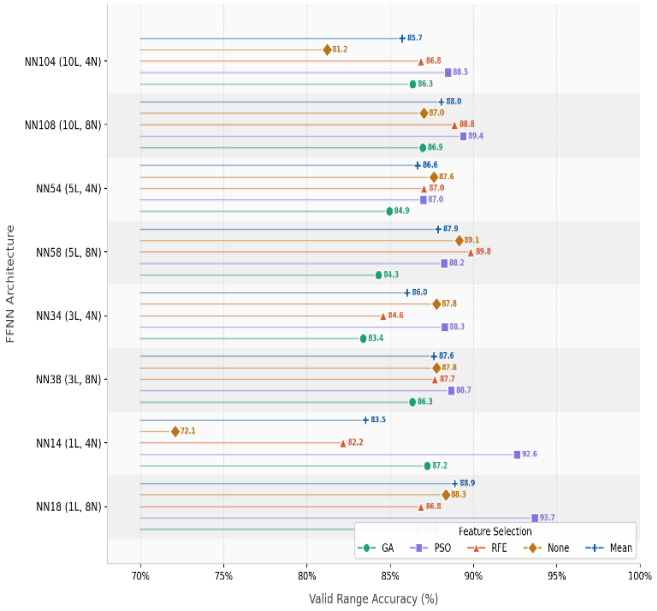


Fig. 5. Valid Range accuracy (%) for FFNN configurations in the Parallel Architecture. L = Hidden Layers; N = Nodes per Layer.

D. Popularity Architecture: Unsupervised Learning Results

The Popularity Architecture — which integrates K-Modes/K-Prototype clustering as the predictive mechanism — achieved the highest coverage across all experimental conditions, as shown in Table V. The no-feature-selection condition produced the best coverage at 96.91%, followed by GA at 95.87%, PSO at 94.10%, and RFE at 92.64%. The coverage range across conditions is narrow (approximately 4%), suggesting that the clustering-based approach is relatively insensitive to the specific feature subset employed — a practically significant property, as it implies that the Popularity Architecture can be deployed without the computational overhead of feature selection.

However, coverage alone does not fully characterize predictive quality. The mean interval width W reveals that the no-feature-selection condition, despite its highest coverage, produces the widest intervals (W = 0.28), shared with PSO. The GA condition, by contrast, achieves a coverage-width efficiency score E = 4.79 — the highest across all configurations — by combining competitive coverage (95.87%) with the narrowest intervals (W = 0.20), a result attributable to the finer cluster granularity of k = 11 compared to k = 5 for the no-selection condition. The M-MAE values (0.040–0.059) confirm that interval midpoints remain close to the true C_avg across all conditions, providing a scalar error measure consistent with standard regression benchmarks.

TABLE V. SUMMARY OF PEAK VALID RANGE ACCURACY RESULTS ACROSS ALL ARCHITECTURES, COMPARED TO THE LITERATURE BASELINE

Architecture	Feature Selection	k (clusters)	Coverage (%)	W (mean width)	M-MAE	Max W (scale)	E = Coverage/W	Diff. vs Baseline (%)
Popularity (UL)	None	5	96.91	0.28	0.053	0.90	3.46	+15.91
Popularity (UL)	GA	11	95.87	0.20	0.040	0.90	4.79	+14.87
Popularity (UL)	PSO	5	94.10	0.28	0.058	0.90	3.36	+13.10
Popularity (UL)	RFE	7	92.64	0.24	0.053	0.90	3.86	+11.64
Parallel (RFC)	None	—	93.51	0.28	0.059	0.90	3.34	+12.51
Parallel (NN18)	PSO	—	93.68	0.26	0.055	0.90	3.60	+12.68
Literature Baseline	—	—	81.00	—	—	—	—	—

The 14%–16% improvement over the supervised baseline warrants careful interpretation. Popularity Architecture benefits from a distinct information structure: rather than learning a mapping from target labels, it groups students by the similarity of their enrollment characteristics and inherits performance bounds from the cluster. This approach is particularly well-suited to the present dataset, where demographic and socioeconomic characteristics appear more informative for group-level stratification than for individual-level regression. The finding is broadly consistent with work by Ding et al. [20], who showed that unsupervised feature representations can boost downstream prediction performance in MOOCs.

E. Overall Analysis and Framework Robustness

Considered holistically, and excluding the anomalous XGBC results, the framework achieves an average accuracy of approximately 86% (range 61%–97%), representing a 5% mean improvement over the 81% baseline (Table V). Including XGBC, the mean drops to 84% (a 3% improvement), with a worst-case accuracy of 31%, reflecting the unsuitability of that specific algorithm for this dataset rather than a systemic weakness of the framework. In both scenarios, the upper-bound accuracy (97%) substantially exceeds the top-end of the baseline range (90%), confirming that the framework can produce state-of-the-art performance even in its worst-case-excluded form.

The robustness of the framework is attributable to the systematic design of its constituent layers. The Anchored Training Data layer reduced training set size from the full 9,300 observations to as few as 3,520 in some conditions, indicating that the algorithm identified smaller, more balanced subsets as preferable training configurations. The in-depth data analysis layer identified key dataset weaknesses—particularly the Lives Environment imbalance and the low dimensionality-reduction fidelity—which informed the decision to restrict the experiment to graduate students with complete enrollment data, thereby maximizing data quality. The multi-algorithm feature selection layer provided evidence that the dataset's predictive power is

broadly distributed across its features, explaining why the no-feature-selection condition frequently performed best (Table V, rows 1 and 5).

To evaluate whether the observed accuracy differences between feature-selection conditions are statistically significant, Wilcoxon signed-rank tests were conducted for each pairwise comparison (No-FS vs GA, No-FS vs PSO, No-FS vs RFE) across the Valid Range accuracy scores of all classifiers. In no comparison did the test yield a statistically significant difference (all $p > 0.05$), confirming that the accuracy differences between feature-selection conditions are not reliably distinguishable from random variation. This finding is consistent with the hypothesis that when features are broadly informative (as in the present dataset), the marginal gain from pruning uninformative variables is insufficient to produce measurable performance improvements [28]. Importantly, the inclusion of three feature-selection paradigms in the framework is not motivated by accuracy improvement claims, but by the methodological goal of systematic cross-paradigm comparison — a contribution that is itself novel in the educational data mining literature, where single-method feature selection remains the norm.

The finding that FFNN depth does not improve accuracy in this context is educationally and technically informative. It implies that the non-linear complexity of the student-performance prediction problem does not exceed what a single-layer network can approximate, given the variable types and sample size involved. This finding discourages unnecessary model complexity in similar institutional settings and provides a principled justification for preferring shallower, faster-training architectures in practice.

VI. Conclusion

This paper presented a novel seven-layer computational framework for predicting student performance ranges in higher education. The framework addresses five identified gaps in the existing literature: the absence of range-valued predictions, insufficient treatment of imbalanced data, limited application of unsupervised learning, shallow data analysis practices, and the lack of systematic neural network architecture comparisons. Validated on a rich dataset of over 200,055 graduate records from the Ibn Zohr University of Morocco, the framework achieved Valid Range accuracy of 84%/86% (worst/common case) with a 3%/5% improvement over a literature baseline of 81%. The unsupervised Popularity Architecture attained peak accuracy of 96.91%, the highest reported in this experimental context.

Five concrete contributions emerge from this work. First, the performance range output format provides a more realistic representation of student academic capability, explicitly acknowledging the influence of life circumstances on grade variability. Second, the Anchored Training Data algorithm offers a practical, computationally inexpensive method for reducing imbalanced data effects without requiring complex oversampling strategies. Third, the Popularity Architecture demonstrates that K-Modes/K-Prototype clustering can function as a direct predictive mechanism for academic performance, opening a new methodological direction in educational data mining. Fourth, the eight-FFNN comparison provides empirical evidence that increasing hidden-layer depth does not improve predictive accuracy in this context, informing future model design choices. Fifth, the comprehensive data analysis layer exposes dataset properties—including socioeconomic attainment gradients and the limitations of dimensionality reduction—that have practical implications for institutional policy.

Several directions merit future investigation. Replicating the framework on datasets from multiple institutions would strengthen external validity and enable cross-institutional comparisons. Enriching the feature space with health, financial, and detailed parental variables could expose additional predictive signals, subject to ethical and legal data-governance constraints. Extending the imbalance-mitigation layer with rule-based sampling architecture, and exploring additional clustering algorithms in the Popularity Architecture, represent tractable methodological improvements. Finally, packaging the framework as an API-accessible microservice would facilitate integration with existing student performance dashboard systems, realizing its practical impact for institutional decision-making.

References

[1] S. W. Pyke et P. Sheridan, « Logistic Regression Analysis of Graduate Student Retention », Can. J. High. Educ., vol. 23, no 2, p. 44-64, août 1993, doi: 10.47678/cjhe.v23i2.183161.

[2] A. I. Adekitan et E. Noma‐Osaghae, « Data mining approach to predicting the performance of first year student in a university using the admission requirements », Educ. Inf. Technol., vol. 24, no 2, p. 1527-1543, déc. 2018, doi: 10.1007/s10639-018-9839-7.

[3] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt, et E. Cascallar, « Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation », Comput. Educ. Artif. Intell., vol. 2, p. 100018-100018, janv. 2021, doi: 10.1016/j.caeai.2021.100018.

[4] H. Waheed et al., « Balancing sequential data to predict students at-risk using adversarial networks », Comput. Electr. Eng., vol. 93, p. 107274-107274, juin 2021, doi: 10.1016/j.compeleceng.2021.107274.

[5] S. Alwarthan, N. Aslam, et I. U. Khan, « Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review », Appl. Comput. Intell. Soft Comput., vol. 2022, p. 1-26, sept. 2022, doi: 10.1155/2022/8924028.

[6] A. Hellas et al., « Predicting academic performance: a systematic literature review », p. 175-199, juill. 2018, doi: 10.1145/3293881.3295783.

[7] M. A. Chaudhry et E. Kazim, « Artificial Intelligence in Education (AIEd): a high-level academic and industry note 2021 », AI Ethics, vol. 2, no 1, p. 157-165, juill. 2021, doi: 10.1007/s43681-021-00074-z.

[8] V. Dignum, « The role and challenges of education for responsible AI », Lond. Rev. Educ., vol. 19, no 1, janv. 2021, doi: 10.14324/lre.19.1.01.

[9] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, et A. Hussain, « Educational data mining to predict students' academic performance: A survey study », Educ. Inf. Technol., vol. 28, p. 905-971, janv. 2023.

[10] B. Alnasyan, M. Basheri, et M. O. Alassafi, « Deep Learning Techniques for Predicting Student's Academic Performance on Virtual Learning Environments: A Review », Research Square (Research Square). Research Square (United States), janvier 2024. doi: 10.21203/rs.3.rs-3888441/v1.

[11] S. J. H. Yang, O. H. T. Lu, A. Y. Q. Huang, J. Huang, H. Ogata, et A. J. Q. Lin, « Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis », J. Inf. Process., vol. 26, p. 170-176, janv. 2018, doi: 10.2197/ipsjjip.26.170.

[12] N. Pereira, « Factor Analysis of Mixed Data (FAMD) and Multiple Linear Regression in R », ARROWDublin Inst. Technol. Dublin Inst. Technol., janv. 2019, [En ligne]. Disponible sur: https://arrow.tudublin.ie/scschcomdis/212

[13] R. R. Utzman, D. L. Riddle, et D. V. Jewell, « Use of Demographic and Quantitative Admissions Data to Predict Performance on the National Physical Therapy Examination », Phys. Ther., vol. 87, no 9, p. 1181-1193, juill. 2007, doi: 10.2522/ptj.20060222.

[14] J. Jaber, N. Keric, P. Kang, et A. Feinstein, « Predicting success: A comparative analysis of student performance on the surgical clerkship and the NBME surgery subject exam », Surg. Open Sci., vol. 1, no 2, p. 86-89, août 2019, doi: 10.1016/j.sopen.2019.07.002.

[15] H. A. Mengash, « Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems », IEEE Access, vol. 8, p. 55462-55470, janv. 2020, doi: 10.1109/access.2020.2981905.

[16] S. D. A. Bujang, A. Selamat, et O. Krejcar, « A Predictive Analytics Model for Students Grade Prediction by Supervised Machine Learning », in IOP Conference Series Materials Science and Engineering, IOP Publishing, févr. 2021, p. 12005-12005. doi: 10.1088/1757-899x/1051/1/012005.

[17] B. J. Kiss, M. Nagy, R. Molontay, et B. Csabay, « Predicting Dropout Using High School and First-semester Academic Achievement Measures », p. 383-389, nov. 2019, doi: 10.1109/iceta48886.2019.9040158.

[18] A. Bilal, « Predicting students' academic performance based on enrolment data », in Proceedings of the 2020 Conference on Applied Machine Learning, janv. 2020, p. 54-61.

[19] N. Z. Zacharis, « Predicting Student Academic Performance in Blended Learning Using Artificial Neural Networks », Int. J. Artif. Intell. Appl., vol. 7, no 5, p. 17-29, sept. 2016, doi: 10.5121/ijaia.2016.7502.

[20] M. Ding, K. Yang, D.-Y. Yeung, et T.-C. Pong, « Effective Feature Learning with Unsupervised Learning for Improving the Predictive Models in Massive Open Online Courses », p. 135-144, févr. 2019, doi: 10.1145/3303772.3303795.

[21] S. Helal et al., « Predicting academic performance by considering student heterogeneity », Knowl.-Based Syst., vol. 161, p. 134-146, juill. 2018, doi: 10.1016/j.knosys.2018.07.042.

[22] Y. Zhao, Q. Xu, M. Chen, et G. M. Weiss, « Predicting Student Performance in a Master's Program in Data Science using Admissions Data. », in Educational Data Mining, janv. 2020. [En ligne]. Disponible sur: https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_27.pdf

[23] B. K. Francis et S. B. Suvanam, « Predicting Academic Performance of Students Using a Hybrid Data Mining Approach », J. Med. Syst., vol. 43, no 6, p. 162-162, avr. 2019, doi: 10.1007/s10916-019-1295-4.

[24] O. E. Aissaoui, Y. E. M. E. Alami, L. Oughdir, A. Dakkak, et Y. E. Allioui, « A Multiple Linear Regression-Based Approach to Predict Student Performance », in Advances in intelligent systems and computing, Springer Nature, 2020, p. 9-23. doi: 10.1007/978-3-030-36653-7_2.

[25] A. Bettahi, F.-Z. Belouadha, et H. Harroud, « A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education », Algorithms, vol. 18, no 10, p. 662-662, oct. 2025, doi: 10.3390/a18100662.

[26] M. Fazil, A. Rísquez, et C. Halpin, « A Novel Deep Learning Model for Student Performance Prediction Using Engagement Data », J. Learn. Anal., vol. 11, no 2, p. 23-41, mai 2024, doi: 10.18608/jla.2024.7985.

[27] B. Strøm, T. Falch, et P. Lujala, « Geographical constraints and educational attainment », RePEc Res. Pap. Econ., sept. 2011.

[28] A. Abdulla, G. Baryannis, et I. Badi, « Weighting the Key Features Affecting Supplier Selection using Machine Learning Techniques », Preprints.org, déc. 2019, doi: 10.20944/preprints201912.0154.v1.

[29] S. Greatorex-Voith et A. Anand, « A data-driven framework for identifying high school students at risk of not graduating on time », in Workshop Proceedings, EDM 2015, janv. 2015.

[30] H. Li, C. Lynch, et T. Barnes, « Early Prediction of Course Grades: Models and Feature Selection », ArXiv Cornell Univ., déc. 2018, doi: 10.48550/arxiv.1812.00843.

[31] S. Alturki, I. Hulpuș, et H. Stuckenschmidt, « Predicting Academic Outcomes: A Survey from 2007 Till 2018 », Technol. Knowl. Learn., vol. 27, no 1, p. 275-307, sept. 2020, doi: 10.1007/s10758-020-09476-0.

[32] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, et R. Nawaz, « Predicting academic performance of students from VLE big data using deep learning models », Comput. Hum. Behav., vol. 104, p. 106189-106189, nov. 2019, doi: 10.1016/j.chb.2019.106189.

Digital Infrastructure Transformation in the Public Sector: Explaining IPv6 Adoption Through the UTAUT Framework

Aydın Koçak[1], Uğur Dagtekin[2], Ahmet Kamil Kabakuş[3]
Department of Management Information Systems, Atatürk University, Erzurum, 25240, Türkiye [1, 3]
Department of Finance, Banking and Insurance, Yozgat Bozok University, Yozgat, 66300, Türkiye [2]

***Abstract*—The Internet, as a product of advanced technological development, has evolved through a dynamic and synergistic process. However, its original architecture was not designed to accommodate such unprecedented growth, resulting in fundamental limitations, particularly in the addressing architecture. The exhaustion of the IPv4 address space has emerged as a major sustainability problem for the Internet. To overcome this limitation, Internet Protocol version 6 (IPv6), which provides a significantly larger address space and additional technical capabilities, was standardized by the Internet Engineering Task Force (IETF) in 1998. Despite its technical superiority, IPv4—standardized in 1981—continues to dominate operational networks, indicating that IPv6 adoption has not yet reached expected levels. This study examines the factors influencing IPv6 adoption in public institutions using the Unified Theory of Acceptance and Use of Technology (UTAUT) framework. Survey data were collected from 456 managerial and technical personnel employed in public institutions in Türkiye. Structural Equation Modeling (SEM) was conducted using SPSS and AMOS software. The findings reveal that facilitating conditions significantly affect both effort expectancy and performance expectancy. Furthermore, effort expectancy and performance expectancy positively influence behavioral intention, which in turn has a direct effect on actual IPv6 usage. These results emphasize the critical role of organizational and structural factors in accelerating IPv6 transition within the public sector. The findings further demonstrate that next-generation network deployment represents not only an engineering challenge but also a socio-technical transformation process shaped by human and organizational factors.**

***Keywords*—IPv6; UTAUT; Structural Equation Modeling (SEM); Technology Acceptance Models; internet and network technologies**

I. Introduction

Digital transformation has become a central priority in modern public administration, requiring not only advanced digital services but also a robust, scalable, and sustainable network infrastructure [1], [2]. At the core of this infrastructure lies the Internet Protocol (IP), which enables global connectivity and underpins all digital communication [3]. Following the near-total depletion of IPv4 address pools at global and regional levels, the transition to IPv6 has evolved from a technical upgrade into a strategic necessity for supporting long-term digital growth [4], [5]. Emerging technologies such as 5G, the Internet of Things (IoT), and smart city applications increasingly depend on the scalability and architectural flexibility offered by IPv6-based networks [6]. Despite the well-documented technical advantages of IPv6, including an almost limitless address space, enhanced security mechanisms, and simplified network management, its global adoption remains inconsistent [7]. This disparity, often termed the IPv6 adoption paradox in the mature adoption [9]. This persistent gap suggests that the barriers to IPv6 transition are not purely technical but are strongly associated with organizational and behavioral factors. Accordingly, large-scale IPv6 deployment should be understood not merely as a networking upgrade, but as a digital infrastructure transformation process requiring both engineering readiness and institutional preparedness. The existing literature on IPv6 has predominantly focused on protocol specifications, performance comparisons, and technical transition mechanisms such as dual-stack and tunneling [10], [11]. While these studies provide valuable engineering insights, they offer limited explanations for why technically mature solutions encounter resistance during organizational implementation. In particular, empirical research addressing the socio-technical dimensions of IPv6 transition—especially from the perspective of IT professionals responsible for planning, deployment, and maintenance in public institutions remains limited. Addressing these behavioral and perceptual barriers is, therefore, critical to understanding the human dimension of digital infrastructure transformation. To address this gap, this study employs the Unified Theory of Acceptance and Use of Technology (UTAUT) framework to examine the determinants of IPv6 adoption in public institutions. Recognizing that infrastructure technologies differ fundamentally from voluntary, user-facing information systems, the UTAUT model is adapted to reflect the institutional context of public-sector infrastructure transitions. Specifically, the model emphasizes facilitating conditions as an antecedent shaping internal cognitive perception—namely, effort expectancy and performance expectancy—while excluding social influence, given the mandatory and policy-driven nature of technology adoption in public organizations. This approach allows infrastructure transition to be analyzed as a socio-technical adoption process rather than a purely technical deployment decision. The contributions of this study are threefold. First, it is among the few empirical investigations examining IPv6 adoption within the public sector of a developing economy. Second, it extends technology acceptance research by demonstrating how facilitating conditions indirectly influence adoption through users' cognitive perceptions in the context of deep infrastructure technologies. Third, it provides

actionable insights for policymakers and IT managers, including the importance of structured training programs and capacity-building initiatives aimed at reducing perceived effort and accelerating the transition to next-generation Inter-net protocols.

The remainder of this study is organized as follows: Section II reviews the theoretical background and related literature. Section III presents the research model and hypotheses. It also outlines the methodology. Section IV reports the empirical findings, and discussion of implications, limitations, and directions for future research. Section V concludes the study.

II. Background

A. Foundations and Evolution of the Internet

Computer networks form the backbone of modern information systems by enabling communication between devices through standardized protocols and layered architectures such as the OSI model [12]. These networks range from local to wide-area structures and support essential operations such as data transmission, routing, and resource sharing. Ethernet technology serves as a widely adopted standard for network infrastructure development across different network scales [13].

The evolution of the Internet has been shaped by early developments such as ARPANET and the adoption of TCP/IP protocols, which enabled interoperability across different network systems [14], [15]. Over time, the expansion of internet usage introduced challenges related to scalability and address management, leading to the development of supporting systems such as DNS [16], [17]. These developments transformed the internet into a global infrastructure supporting communication, information exchange, and digital services [18].

B. Internet Protocol and Limitations of IPv4

The Internet Protocol (IP) is the main network-layer protocol that addresses uniquely devices on the internet and allows routing of data packets between networks [2]. Its primary functions are logical addressing, packet encapsulation, routing, and fragmentation. IPv4 uses a 32-bit addressing scheme which translates theoretically to around 4.3 billion unique addresses. However, with the fast growth in mobile devices, cloud computing services and Internet of Things (IoT) applications, this address space has become inadequate. The last allocation by IANA of IPv4 address blocks to Regional Internet Registries in 2011 was a major step as it indicated that IPv4 addresses were exhausted worldwide [19]. Network Address Translation (NAT) and Carrier-Grade NAT (CGNAT) have been used as short-term solutions; however, they do not support the end-to-end connectivity principle, increase the complexity of the network, and bring performance and security problems according to studies conducted by Donley et al. [6] and Richter et al. [7]. These limitations clearly indicate that IPv4 can no longer fulfill modern internet requirements [8].

C. IPv6 and the Next-Generation Internet Architecture

IPv6 was developed by the Internet Engineering Task Force to provide a solution for the limitations in addressing and architecture that IPv4 presented [3]. It is structured with 128 bits for its address space which means theoretically it can support about 3.4×1038 addresses, hence solving any problems related to exhaustion of addresses permanently [20]. Some important technical benefits of IPv6 are a streamlined packet header, native support for IPsec, SLAAC, improved multicast and anycast features, and removal of the requirement for NAT [21], [9]. All these make room for more secure networks with better scalability and manageability, especially in large public sector network infrastructures. However, despite these technical advantages, the success of IPv6 deployment depends not only on protocol capabilities but also on organizational capacity, user readiness, and institutional support mechanisms.

D. Technology Adoption and the UTAUT Model

The adoption of technological innovations is a process that takes place within the framework of several dimensions, namely individual perceptions, organizational structures, and social interactions. Theoretical models developed to explain this process include the Theory of Reasoned Action (TRA), the Theory of Planned Behavior (TPB), Innovation Diffusion Theory (IDT), and Technology Acceptance Model (TAM) [22], [23], [24]. These models were developed independently and each address only part of the adoption process. The Unified Theory of Acceptance and Use of Technology (UTAUT) was introduced by Venkatesh et al. [11] as a synthesis of eight previous models, providing a high level of theoretical integration for technology acceptance and use. The model explains behavioral intention and actual use through four constructs: performance expectancy, effort expectancy, social influence, and facilitating conditions. UTAUT has been chosen as the framework for analysis in this study on IPv6 transition because it explains about 70% of variance, includes moderator variables such as age, gender, experience, and voluntariness of use, as well as organizational factors. By applying UTAUT to a foundational network technology, this study extends technology acceptance research beyond user-facing systems to deep digital infrastructure environments.

III. Research Model and Hypotheses

In this study, the Unified Theory of Acceptance and Use of Technology (UTAUT) is adopted as the theoretical framework to identify the factors influencing the adoption of IPv6 technology in public institutions in Turkey. UTAUT provides a comprehensive approach that is widely used to explain individuals' perceptions and intentions toward technology use [11]. Within the scope of this study, the research model focuses on facilitating conditions, effort expectancy, performance expectancy, behavioral intention, and actual use behavior. The social influence construct was not included, as IPv6 adoption in public institutions is predominantly driven by formal mandates and organizational policies rather than peer-based or voluntary social dynamics. Confirmatory Factor Analysis (CFA) and Structural Equation Modeling (SEM) were conducted using SPSS AMOS. Unlike voluntary technology adoption contexts, IPv6 implementation in public institutions is largely driven by regulatory frameworks and institutional directives. Therefore, the social influence construct, which reflects interpersonal or peer effects, is considered less relevant in this context. Similarly, voluntariness of use is inherently limited, as employees are expected to comply with organizational technology policies. For this reason, the UTAUT model was conceptually adapted by

excluding social influence and related moderating variables, allowing the analysis to focus on organizational and cognitive determinants of adoption. Social influence does not always emerge as a significant determinant even though it is part of UTAUT framework when dealing with highly specialized contexts such as institutional technical infrastructure transitions. By excluding this particular dimension, conceptual focus increases while improving model fit in an organizational context with this specific sample size. This modeling decision agrees with findings reported in similar contexts within literature [25].

A. Expected Contributions of the Study

The results of this research are expected to enhance understanding of the transition processes to IPv6 in public institutions and help identify and ameliorate those factors that delay or impede such a transition. Since IPv6 is an important component of infrastructure for the digital transformation of public institutions, its accelerated adoption can be facilitated by empirical evidence from managerial and technical personnel perspectives. The results should indicate organizational and technical areas that need improvement for a better IPv6 transition process. The UTAUT model used in this study assesses how much the behavioral intention and actual usage behavior are determined by performance expectancy, effort expectancy, and facilitating conditions. Therefore, it attempts to get specific information about priority areas for public institutions concerning IPv6 adoption. In this sense, this study not only informs about the approaches taken by public institutions toward IPv6 but also helps in strategic decision-making concerning wider technological as well as digital transformation efforts. In addition, it attempts to be an original contribution to literature where very few empirical studies have been undertaken regarding the transition from IPv4 to IPv6 in the course of digital transformation. The proposed structural model is also anticipated as one of the key outcomes of this research to be a reference model for future studies carried out under similar organizational contexts. Unlike prior IPv6 studies focusing on protocol performance or transition mechanisms, this research conceptualizes IPv6 migration as an organizational and behavioral transformation process within public digital infrastructure.

B. Measurement Instruments

All constructs in this study were measured using validated scales adapted from prior UTAUT literature. The survey instrument consists of items measuring performance expectancy, effort expectancy, facilitating conditions, behavioral intention, and actual use. The measurement structure of the study, including constructs, item codes, and sources, is presented in Table I. The factor structure of the scales has been previously validated in the literature; therefore, the measurement model was assessed using confirmatory factor analysis (CFA) to ensure construct validity and consistency. All items were measured using a five-point Likert scale (1 = Strongly disagree, 5 = Strongly agree).

TABLE I. MEASUREMENT CONSTRUCTS AND ITEMS

Construct	Code	Measurement Item	Source
Performance Expectancy (PE)	PE1, PE2, PE3, PE4	Improves task efficiency; enhances organizational performance; strengthens cybersecurity; supports effective process management.	Venkatesh et al. (2003)
Effort Expectancy (EE)	EE1, EE2, EE3, EE4	Easy to learn and use; transition process is simple and understandable; requires little effort; less complex than existing systems.	Venkatesh et al. (2003)
Facilitating Conditions (FC)	FC1, FC2, FC3, FC4	Infrastructure support available; sufficient technical knowledge; technical support accessible; training and guidance provided.	Venkatesh et al. (2003)
Behavioral Intention (BI)	BI1, BI2, BI3, BI4	Intend to use IPv6; willing to participate in transition; committed to using IPv6; support its future adoption.	Venkatesh et al. (2003)
Actual Use (AU)	AU1, AU2, AU3	Regular usage; increasing adoption; integrated into daily processes.	Venkatesh et al. (2003)

C. Research Model

The conceptual model developed in this study is intended to assess the IPv6 adoption process in public institutions within the framework of the Unified Theory of Acceptance and Use of Technology (UTAUT). This model comprises five core constructs—Performance Expectancy, Effort Expectancy, Facilitating Conditions, Behavioral Intention, and Actual Use Behavior—designed to explicate employees' perceptions and usage behavior regarding IPv6 technology. These constructs together represent individual cognitive evaluations as well as organizational and technical conditions through an integrated approach. In the model, Performance Expectancy represents employee perceptions that IPv6 technology will improve organizational efficiency, network performance, and business processes. Effort Expectancy refers to perceived ease of learning and using IPv6 while Facilitating Conditions represent institutional technical infrastructure, training opportunities, technical support mechanisms, and overall organization readiness. These variables are assumed to have a direct effect on employees' Behavioral Intention toward using IPv6 technology. Behavioral Intention is what reflects individuals' willingness to use IPv6; it serves as the main determinant of Actual Use Behavior in this model. Furthermore, Facilitating Conditions are assumed to have a direct effect on actual usage independent of behavioral intention. Besides that, following previous studies based on UTAUT, Effort Expectancy is hypothesized to positively influence Performance Expectancy; this linear relation is explicitly specified in the model. Prior to data collection, approximately 600–700 IT personnel working in public institutions were contacted. The final sample consists of 456 respondents, comprising managerial and technical IT personnel employed in state universities in Türkiye. Survey items were adapted from measurement scales used in prior studies. This study adopted a quantitative research design using the survey method for data collection. The survey was administered between February 1, 2025, and July 24, 2025. Participants were selected based on convenience sampling method of accessibility. To ease participation, the survey was conducted online via Google Forms. The data collection instrument that was developed to test the research model consists of a total of 26 items. Besides seven questions capturing

demographic characteristics of participants, four items each for Performance Expectancy measurement, Effort Expectancy measurement, Facilitating Conditions measurement, Behavioral Intention measurement, and three items measuring Actual Use Behavior were included in the survey. The measurement items were based on scales widely used in UTAUT literature. The theoretical factors affecting the transition process of IPv6 in public institutions are described by the conceptual research model developed herewith. This model serves as an analytical basis for testing the hypotheses proposed in this study. The research model can be seen in Fig. 1.

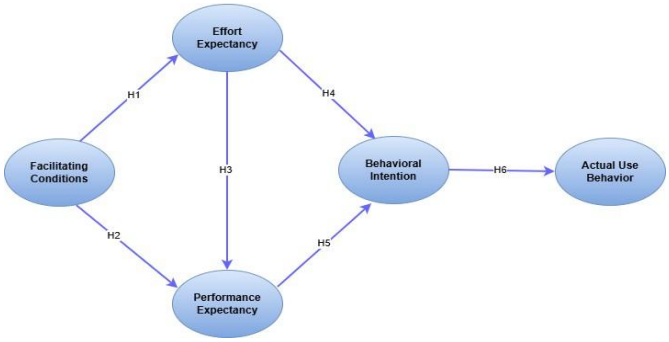


Fig. 1. Our research model.

The measurement scales used in this study were adapted from previously validated instruments in the literature. As presented in Table II, all constructs in the research model—including performance expectancy, effort expectancy, facilitating conditions, behavioral intention, and actual use behavior—were measured using multi-item scales. Demographic information was collected through seven items, while the remaining constructs were measured using between three and four items each. All scale items were derived from the Unified Theory of Acceptance and Use of Technology (UTAUT) framework and adapted to the context of IPv6 adoption in public institutions [26].

TABLE II. MEASUREMENT SCALES USED IN THE RESEARCH MODEL AND THEIR SOURCES.

Scales	Number of Items per Construct	Source
Demographic Information	7	[26]
Performance Expectancy	4	[26]
Effort Expectancy	4	[26]
Facilitating Conditions	4	[26]
Behavioral Intention	4	[26]
Actual Use Behavior	3	[26]

D. Hypotheses

The model primarily aims to test the hypotheses articulated below and to analyze the results obtained.

H1: Facilitating conditions have a positive and significant impact on effort expectancy in learning and using IPv6 technology by employees of public institutions.

H2: Facilitating conditions have a positive and significant impact on performance expectancy as perceived by employees in terms of improving organizational performance with IPv6 technology.

H3: Effort expectancy has a positive and significant impact on performance expectancy as perceived by employees in terms of improving organizational performance with IPv6 technology.

H4: Effort expectancy has a positive and significant impact on the behavioral intention of employees toward using IPv6 technology.

H5: Performance expectancy has a positive and significant impact on the behavioral intention of employees toward using IPv6 technology.

H6: Behavioral intention has a positive and significant effect on the actual usage behavior of IPv6 technology in public institutions.

E. Data Analysis Methods

The data analysis was done in a specific order, taking into account how each statistical step influenced the following analyses. The first step included data cleaning, which involved checking for missing values, outliers, and any possible errors in data entry. Normality tests were then applied to determine if the data met the requirements for analysis. The multicollinearity among variables was checked through multiple linear correlation analysis. Reliability analysis checked the internal consistency of the scales, and validity analysis tested the measurement constructs. Finally, correlation analysis was performed to see how different variables relate to each other. In this study, independent variables included demographic information, performance expectancy, effort expectancy, facilitating conditions, and behavioral intention; however, the dependent variable was actual usage behavior. Data analyses were carried out with SPSS 20.0 and AMOS 24.0 software packages. The proposed hypotheses were tested using Structural Equation Modeling (SEM). The results from these analyses are discussed under appropriate subheadings within the context of the research model.

F. Reliability Analysis

Reliability analysis is a statistical approach that assesses the internal consistency among scale items and gauges how well a scale captures the true essence of what it intends to measure, both accurately and consistently. The main focus was on Cronbach's alpha coefficient, which varies between 0 and 1. As the alpha value approaches 1, this indicates stronger internal consistency for the scale. The ranges of alpha coefficients interpreted in this research were based on definitions by other authors [27]–[29] and are presented below: A coefficient within the interval 0.00 ¡ 0.40 implies that the scale is unreliable. If it falls within the range of 0.40 ¡ 0.60, then there is low reliability for that scale. A moderate reliability would be indicated by a coefficient from 0.60 up to but not including 0.80, while high reliability would be signified when the value is above 0.80 up to 1.00. Table III displays Cronbach's alpha values for factors and their respective sub-items as utilized in this study.

TABLE III. CRONBACH'S ALPHA COEFFICIENTS FOR THE FACTORS

Factor	Cronbach's Alpha
Performance Expectancy	0.902
Effort Expectancy	0.881
Facilitating Conditions	0.770
Behavioral Intention	0.875
Actual Use Behavior	0.845

The coefficients of Cronbach's alpha computed for the factors and their respective items reveal a high degree of reliability, which is indicative of very strong internal consistency of the scale. It can be seen from Table III that Performance Expectancy has the highest internal consistency (= 0.902) and Facilitating Conditions has the lowest consistency (= 0.770).

G. Validity Analysis

Validity analysis is the determination of whether a measuring instrument or test truly and appropriately measures the construct that it claims to measure [30], [31]. Validity makes research results meaningful, relevant, and credible by permitting correct inferences to be made and sound decisions based on the trustworthiness of the results [32].

H. Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) is an essential analysis technique for the evaluation of the construct validity of scales and questionnaires. After the theoretically proposed factor structure has been specified, CFA will indicate how well the observed data fit this structure. It assesses factor loadings, factor correlations, and model parameters to determine the extent to which a proposed model aligns with observed data. CFA, thus, is very important in determining whether or not a measurement instrument adequately and accurately measures its intended constructs [33, p. 2]. In this work, CFA was performed for construct validity testing. The factor structure obtained is illustrated in Fig. 2.

The CFA showed that all factor loadings were greater than 0.5, thus confirming their adequacy [34], and the model fit indices met the recommended thresholds. No misfits were detected, and consequently, no items or factors needed to be removed. Table IV provides a summary of the definitions of the fit indices, their acceptable ranges, and the values that resulted from this analysis.

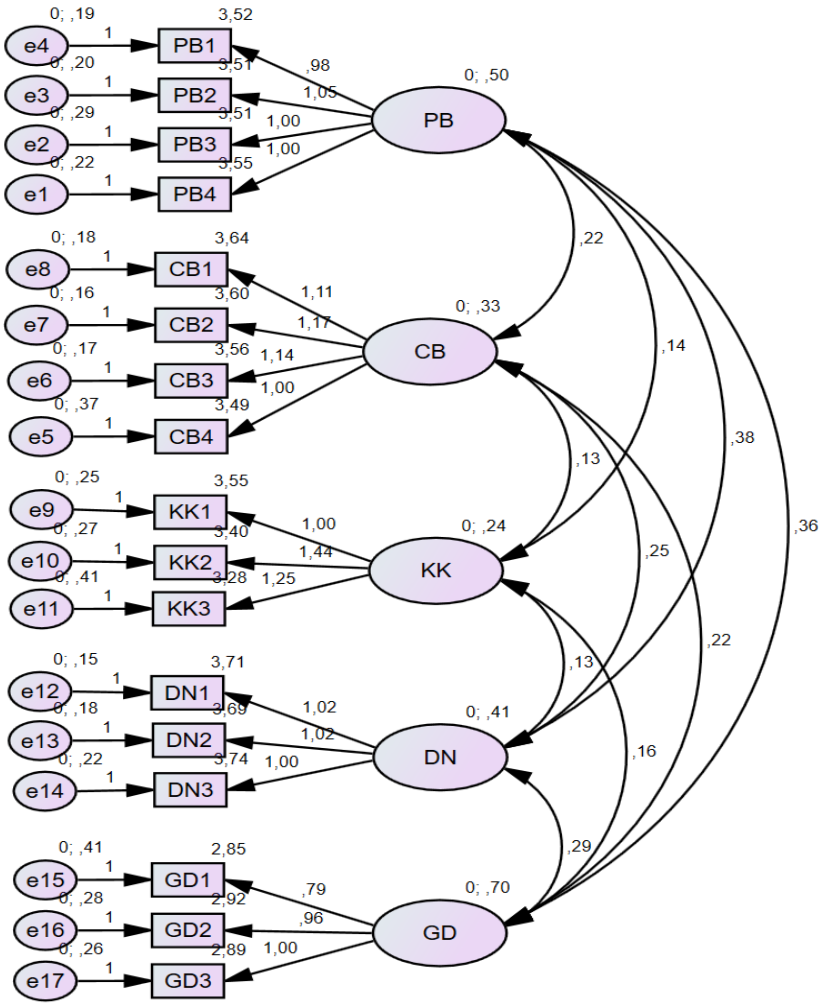


Fig. 2. Our SPSS AMOS CFA (Confirmative Factor Analysis). Source: Author's own illustration generated in SPSS AMOS.

TABLE IV. MODEL FIT INDICES

Factor	Description	Acceptable Fit	Cronbach's Alpha
CMIN/DF	Reduces sample sensitivity and reflects the overall fit of the model to the data in a more balanced way.	$2.00 \leq CMIN/DF \leq 5.00$	2.607
CFI	Assesses the fit by comparing the model with an independent model.	$0.90 \leq CFI \leq 0.95$	0.963
RMSEA	Indicates how closely the model approximates the actual data.	$0.05 \leq RMSEA \leq 0.10$	0.059
PClose	It is the standardized measure of the differences between the model and the observed data.	$0.05 \leq SRMR \leq 0.10$	0.038

The Confirmatory Factor Analysis (CFA) results show that the model has a good fit. The CMIN/DF value was calculated as 2.607, the CFI value as 0.963, the RMSEA value as 0.059, and the PClose value as 0.033. All these indices fall within the acceptable ranges recommended in the literature [35], [36].

These results indicate that the proposed measurement model fits the observed data very well and confirms the validity of the construct structure. Factor loadings, R^2 values, AVE, CR, and MSV are given in Table V as indicators for assessing reliability and validity for constructs being measured.

TABLE V. VALIDITY VALUES OF THE SCALE ITEMS

Factor	Item	Factor Weight	R^2	AVE	CR	MSV
Performance Expectancy	PB1	0.848	0.719	0.679	0.902	0.694
	PB2	0.857	0.735			
	PB3	0.796	0.633			
	PB4	0.837	0.700			
Effort Expectancy	CB1	0.829	0.688	0.656	0.883	0.473
	CB2	0.863	0.745			
	CB3	0.845	0.714			
	CB4	0.690	0.476			
Facilitating Conditions	KK1	0.696	0.485	0.534	0.774	0.226
	KK2	0.801	0.642			
	KK3	0.690	0.476			
Behavioral Intention	DN1	0.720	0.736	0.700	0.875	0.694
	DN2	0.840	0.705			
	DN3	0.811	0.658			
Actual Use Behavior	GD1	0.720	0.518	0.649	0.846	0.367
	GD2	0.834	0.696			
	GD3	0.856	0.732			

Factor loadings were between 0.690 and 0.863 and all were above the minimum threshold of 0.50, which indicates that item reliability is strong. Convergent and discriminant validity are assessed with the help of the coefficient of determination (R^2), Average Variance Extracted (AVE), Composite Reliability (CR), and Maximum Shared Variance (MSV). Values for all factors about their items are presented in Table V; results here confirm that high reliability and construct validity exist across all constructs for the measurement model.

I. Factor Loadings and Item-Explained Variance

All item factor loadings are between 0.690 and 0.857 as seen in the table. In the structural equation modeling literature, factor loadings that exceed 0.50 suggest that items sufficiently represent their respective latent constructs [37], [38]. Therefore, it can be concluded that all items have acceptable loadings. The R^2 values indicate that the percentage of variance explained by each item ranges from 0.476 to 0.745. This finding confirms that a considerable part of the variance for each item is explained by its related latent variable, which proves sufficient measurement power [39].

J. Convergent Validity

Convergent validity was evaluated via Average Variance Extracted (AVE) values. A value of AVE greater than 0.50 is documented in the literature as indicating that more than half the variance in the indicators is accounted for by the latent construct, hence establishing convergent validity [40]. In this particular study:

- Performance Expectancy held an AVE of 0.679;
- Effort Expectancy has an AVE of 0.656;
- Facilitating Conditions bear an AVE equal to or above 0.534;
- Behavioral Intention carries an AVE of 0.700;
- Actual Use has an AVE at 0.649.

All these are above the 0.50 cut-off point and hence, convergent validity is established for all the constructs.

K. Internal Consistency and Composite Reliability

The evaluation of the reliability of the scales was done through Composite Reliability. A CR value greater than 0.70 is an indication of adequate internal consistency of the construct as per [37]. The results from the analysis indicated that the CR values fell within a range from 0.774 to 0.902, which further confirmed that all constructs had high reliability.

L. Discriminant Validity

Discriminant validity was evaluated based on the Maximum Shared Variance (MSV) criterion. According to Fornell and Larcker's guideline, discriminant validity is established if the MSV of a construct is lower than its Average Variance Extracted (AVE) [40]. In this study, all constructs satisfy the condition MSV ¡ AVE, thus proving the statistical distinctiveness of the constructs and confirming that discriminant validity is achieved.

M. Overall Evaluation

Based on the results: All items have acceptable and strong factor loadings. The measurement model has convergent validity (AVE). Constructs exhibit high internal consistency and reliability (CR). Discriminant validity (MSV ¡ AVE) is confirmed for all constructs. In brief, the measurement model

applied in this study fulfills all SEM literature requirements concerning validity and reliability.

N. Discriminant Validity Assessment

Discriminant validity is one of the ways to test the model's validity. It proves that each construct is unique by showing that they are not related or dependent on each other [41]. To assess this, both Fornell–Larcker and HTMT analyses were performed. The results from the Fornell–Larcker test are shown in Table VI.

TABLE VI. FORNELL-LARCKER TEST

Factors	PB	CB	KK	DN	GD
Performance Expectancy	**0.835**				
Effort Expectancy	0.546	**0.810**			
Facilitating Conditions	0.411	0.476	**0.731**		
Behavioral Intention	0.833	0.687	0.421	**0.837**	
Actual Use Behavior	0.606	0.458	0.387	0.539	**0.805**

The square root of the average variance extracted for each construct should, according to Fornell and Larcker, be greater than the correlation coefficients of that construct with other constructs [40]. Results indicated that all constructs satisfy this condition; therefore, discriminant validity is achieved. HTMT analysis is another approach used to check discriminant validity. The output from the HTMT analysis will be shown in Table VII below:

TABLE VII. HTMT TEST

Factor	PB	ÇB	KK	DN	GD
Performance Expectancy					
Effort Expectancy	0.545				
Facilitating Conditions	0.418	0.478			
Behavioral Intention	0.832	0.686	0.427		
Actual Use Behavior	0.594	0.478	0.416	0.525	

Henseler et al. developed the HTMT criterion, which states that if values are lower than 0.85 or 0.90, then discriminant validity is achieved [42]. A review of the results reveals that all values fall under 0.85; thus, this confirms the existence of discriminant validity.

O. Correlation Analysis

Correlation analysis is a statistical method for assessing the linear relationship between two continuous variables. The strength and direction of the linear relationship are defined by the correlation coefficient, which varies from -1 to +1. A positive coefficient indicates a direct relationship between variables, whereas a negative coefficient indicates an inverse relationship.

The closer the coefficient is to zero, the weaker the relationship; as it approaches ±1, the relationship becomes increasingly linear [43]. Correlation analysis was therefore performed in this study to assess how these factors relate with one another and whether such relationships are significant or not. Results are shown in Table VIII.

TABLE VIII. FACTOR CORRELATION ANALYSIS

Factor	PB	CB	KK	DN	GD
Performance Expectancy	1				
Effort Expectancy	0.546	1			
Facilitating Conditions	0.411	0.476	1		
Behavioral Intention	0.833	0.687	0.421	1	
Actual Use Behavior	0.606	0.458	0.387	0.539	1

P. Correlation Analysis Among Factors

In this study, the relationships between latent variables were tested using the correlation coefficients derived from confirmatory factor analysis. The size and sign of the correlation coefficients were assessed in order to know about the level of linear relationships existing among these factors. In SEM and behavioral sciences, the following threshold values are generally used for interpreting correlation coefficients: 0.10–0.29 low; 0.30–0.49 moderate; 0.50–0.69 moderate-high; and 0.70 and above high correlation [44], [37]. The analysis results indicated that correlation coefficients were above 0.40, which means that there is a statistically significant and moderately strong linear relationship between the factors. The range of values taken by these correlation coefficients between factors fell from 0.387 to 0.833. There is a moderate-high positive relationship between Performance Expectancy and Effort Expectancy (r = 0.546). Performance Expectancy has a high-level relationship with Behavioral Intention (r = 0.833). There is a moderate-high positive relationship between Effort Expectancy and Behavioral Intention (r = 0.687). The relationships of Facilitating Conditions with other factors are at a moderate positive level (r = 0.387–0.476), which indicates that this construct creates meaningful but not very strong connections with the other variables. The correlation coefficient between Behavioral Intention and Actual Use is moderate-high positive (r = 0.539). All the correlation coefficients are positive and above 0.30, which means that statistically significant and theoretically consistent linear relationships exist among the factors [38].

Q. Structural Equation Modeling

Structural Equation Modeling (SEM) is a common method for estimating parameters and testing hypotheses. It serves theory development by explaining variance through linear equations comprising observed and latent variables. SEM facilitates the simultaneous assessment of measurement models and theoretical constructs. It also allows the integrated evaluation of the measurement model, which includes analyses of reliability and validity, with the structural model used to test hypotheses. In this sense, SEM constitutes a robust statistical tool that helps to test holistically any relationships between independent, mediating, moderating, and dependent variables so as to further develop theoretical models [45]. Therefore, SEM was used to analyze direct as well as indirect relationships among the variables stipulated above. The model resulting from this analysis was then interpreted in order to assess both research hypotheses and the validity of the theoretical framework. The structural model generated from this analysis is shown below in Fig. 3.

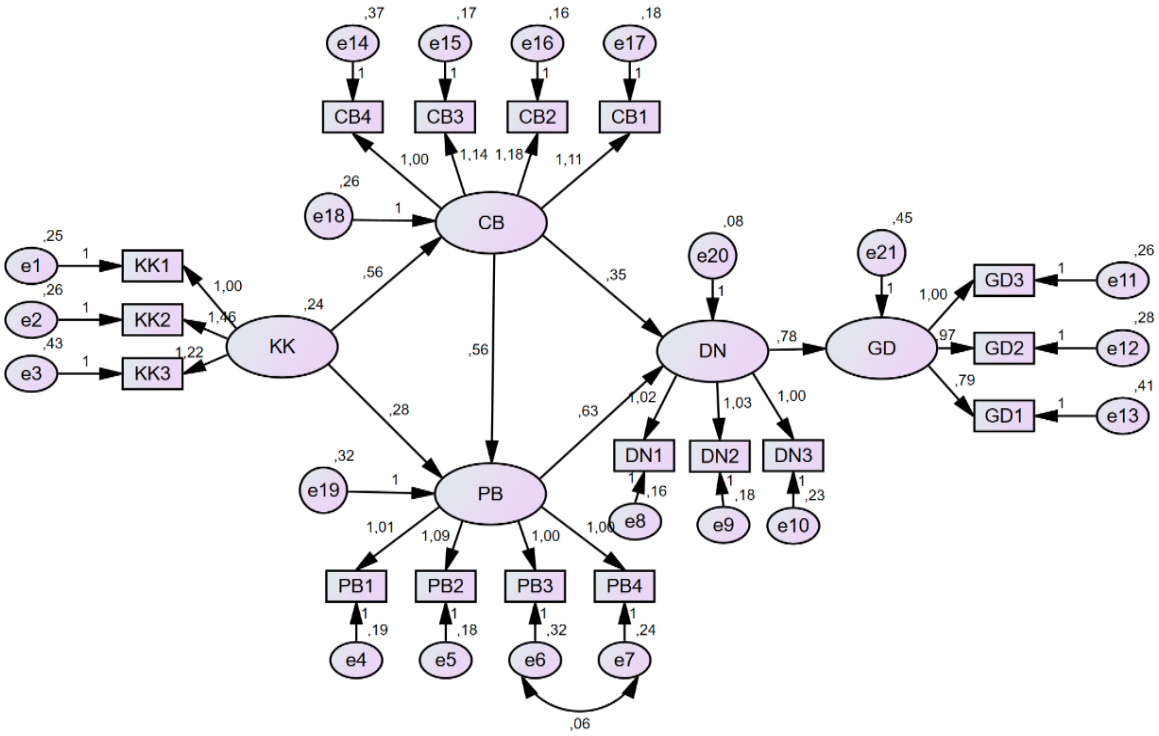


Fig. 3. Structural model. Source: Author's own illustration on SPSS AMOS.

Several fit indices are considered after Structural Equation Modeling (SEM). The acceptable ranges for these indices and the fit values obtained from the analysis are presented in Table IX.

TABLE IX. MODEL FIT EVALUATION

Fit Index	Ideal Fit	Acceptable Fit	Model Fit
CMIN/DF	0.00 ≤ CMIN/DF ≤ 1.00	3.00 ≤ CMIN/DF ≤ 5.00	2.713
CFI	0.95 ≤ CFI ≤ 1.00	0.90 ≤ CFI ≤ 0.95	0.960
RMSEA	0.00 ≤ RMSEA ≤ 0.06	0.05 ≤ RMSEA ≤ 0.6	0.059
PClose	0.00 ≤ SRMR ≤ 0.05	0.05 ≤ SRMR ≤ 0.10	0.038

The results from the Structural Equation Modeling (SEM) analysis are discussed in this section. Model fit indices were found to be CMIN/DF = 2.713, CFI = 0.960, RMSEA =0.059, and SRMR = 0.038. Based on the literature, these values indicate that the model has an acceptable and good fit [37], [38]. The CFI and SRMR measurements particularly fall within the ideal fit range, thus proving strong alignment of this particular model with data. An RMSEA measurement that is less than 0.06 means insensitivity to error for this model as well as consistent testing of structural relationships [46].

In view of such results, one may deduce that the formulated structural equation model is congruent with a theoretical framework and fits well with empirical data such that parameters can be interpreted reliably. The fit indices for this model are in agreement with standards from academic literature as well as context-specific adequacy criteria for field studies. Table X below presents standardized regression weights used to assess interactions and relationship strength among factors in the model:

TABLE X. STANDARDIZED REGRESSION WEIGHTS

Factor Relationships			Standardized Regression Weight
Facilitating Conditions	→	Effort Expectancy	0.475
Facilitating Conditions	→	Performance Expectancy	0.194
Effort Expectancy	→	Performance Expectancy	0.464
Effort Expectancy	→	Behavioral Intention	0.770
Performance Expectancy	→	Behavioral Intention	0.681
Behavioral Intention	→	Actual Use Behavior	0.592

Standardized regression weights are coefficients derived from the transformation of variables such that their standard deviations equal one. In other words, these coefficients reveal how many standard deviations the dependent variable will change for a one-standard-deviation increase in the independent variable [47]. Hence, standardized regression weights can be very informative about both the direction and strength of the relationships among variables. Effect sizes are described as small, medium, and large with ranges of 0.10–0.29, 0.30–0.49, and from 0.50 upwards [44]. Looking at Table X, results reveal that Effort Expectancy has a medium effect on Performance Expectancy (= 0.464).

Facilitating Conditions has a weak effect on Performance Expectancy (= 0.194), while it has a medium effect on Effort Expectancy (= 0.475). In addition, Performance Expectancy has a strong effect on Behavioral Intention (= 0.681), and Effort Expectancy has an even stronger effect on Behavioral Intention with = 0.770, indicating that perceived ease of use is decisive in

forming the intention to adopt IPv6 technology. The effect of Behavioral Intention on Actual Use is high with being equal to 0.592. The relationships in the model are mostly at medium and high effect levels with Effort Expectancy and Performance Expectancy being strong determinants of Behavioral Intention, thus proving that the proposed structural model fits theoretical expectations while causal relationships among variables are statistically significant and robust. Moreover, standardized regression weights in this study are consistent with literature, as shown in Fig. 4.

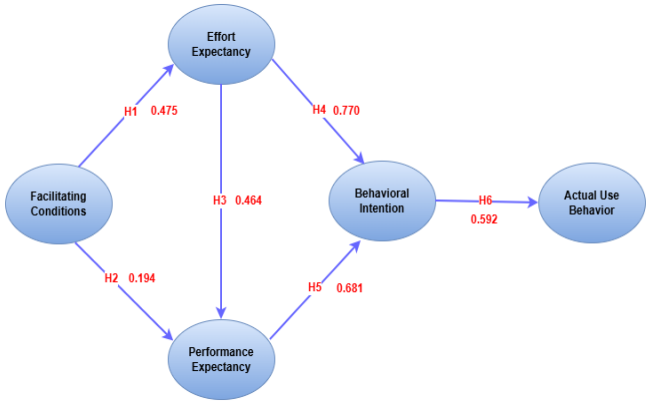


Fig. 4. Standardized regression weights and directions. Source: Author's own illustration.

The standardized regression weights from the model indicate that the direction and magnitude of the relationships among variables are different. Thus, Effort Expectancy has a medium positive effect on Performance Expectancy (= 0.464). This result shows that when learning and using IPv6 technology is perceived as easy, it significantly supports any performance gains expected from the technology. Facilitating Conditions has a weak effect on Performance Expectancy (= 0.194) but, on the other hand, Facilitating Conditions has a medium positive effect on Effort Expectancy (= 0.475). This means that existing technical infrastructure and support mechanisms in the organization help in learning and using IPv6 technology. The effect of Performance Expectancy on Behavioral Intention is high and positive (= 0.681). In addition, the effect of Effort Expectancy on Behavioral Intention is very high (= 0.770). This finding shows that perceived ease of use is a stronger determinant of participants' intention to use IPv6 technology than performance expectancy. Finally, Behavioral Intention has a high positive effect on Actual Use (= 0.592).

Such results that give evidence that good intentions toward IPv6 technology will almost always translate into actual usage behavior. It can be said in general terms that most of the relationships tested in this model have medium to large effect sizes, with Effort Expectancy and Performance Expectancy being significant antecedents to Behavioral Intention. This finding also attests to the fact that the model specified is consistent with the theoretical framework, and all causal relationships among variables are statistically significant and explanatory.

R. Hypothesis Testing

To test the hypotheses formulated in the study, the relationships in the structural model were evaluated. For each path coefficient, the standardized estimate, standard error, critical ratio, and significance level were assessed. Based on these results, decisions were made about whether to accept or reject the hypotheses. This is an approach that determines if the relationships predicted by a model are statistically supported. Results related to these values are given in Table XI.

TABLE XI. EVALUATION OF HYPOTHESES

Hypothesis	Recitations			β	SE	t (C.R.)	p	Hypothesis Results
H1	Facilitating Conditions	→	Effort Expectancy	0.560	0.074	7.584	***	Accept
H2	Facilitating Conditions	→	Performance Expectancy	0.275	0.081	3.381	***	Accept
H3	Effort Expectancy	→	Performance Expectancy	0.558	0.072	7.746	***	Accept
H4	Effort Expectancy	→	Behavioral Intention	0.354	0.048	7.296	***	Accept
H5	Performance Expectancy	→	Behavioral Intention	0.625	0.047	13.234	***	Accept
H6	Behavioral Intention	→	Actual Use Behavior	0.780	0.069	11.250	***	Accept

*p<0.001, *p<0.05

S. Hypothesis Testing Results

The significance of relationships among variables in hypothesis testing was determined using Structural Equation Modeling (SEM). Standardized regression coefficients () were used to indicate effect sizes, where values between 0.10 and 0.29 are considered weak, from 0.30 to 0.49 moderate, and above 0.50 strong [48], [49]. Standard error (S.E.) is a measure of the precision of the estimate, while the Critical Ratio (C.R.) is equivalent to a t-statistic, with —C.R.—greater than 1.96 indicating statistical significance at p ¡.

As shown in Table XI, the results are statistically significant ($p < 0.05$) [50]. Facilitating Conditions have a strong positive effect on Effort Expectancy ($\beta = 0.560$; C.R. = 7.584; $p < 0.001$, H1) and a weaker positive effect on Performance Expectancy ($\beta = 0.275$; C.R. = 3.381; $p < 0.001$, H2). Effort Expectancy has a strong influence on Performance Expectancy ($\beta = 0.558$; C.R. = 7.746; $p < 0.001$, H3) and a moderate effect on Behavioral Intention ($\beta = 0.354$; C.R. = 7.296; $p < 0.001$, H4). Performance Expectancy has a strong effect on Behavioral Intention ($\beta = 0.625$; C.R. = 13.234; $p < 0.001$, H5), and Behavioral Intention strongly predicts Actual Use ($\beta = 0.780$; C.R. = 11.250; $p < 0.001$, H6). These results support the UTAUT model, indicating that Facilitating Conditions influence Effort and Performance Expectancies, which in turn shape Behavioral Intention and lead to the Actual Use of IPv6. The findings emphasize the combined role of individual perceptions and organizational support in achieving high explanatory power.

IV. Discussion

The transition to IPv6 should be considered not merely a technical protocol upgrade, but a complex organizational transformation involving behavioral and structural changes in the public sector. Using a modified UTAUT framework, this study examined the integration of IPv6 adoption among IT professionals in the public sector. The empirical findings provide important insights into the socio-technical dynamics of this major digital infrastructure transfer.

A. Theoretical Implications

One of the key observations emerging from this study concerns the organizational context in which infrastructure technologies such as IPv6 are implemented. In public sector environments, large-scale technological transitions are typically shaped by institutional strategies, regulatory frameworks, and formal organizational directives. Unlike many voluntary technology adoption settings, the implementation of critical infrastructure technologies often follows structured planning processes and strategic roadmaps defined at organizational or national levels.

In contrast to core UTAUT constructs, social influence was not included in the final model due to its limited relevance in the public-sector context examined in this study. In such environments, technology adoption decisions are predominantly driven by formal mandates, regulatory requirements, and institutional directives rather than peer or interpersonal pressure. Preliminary testing indicated that including social influence did not improve model fit, further supporting its exclusion. This finding suggests that, for large-scale infrastructure transitions such as IPv6, organizational and structural factors outweigh social dynamics in shaping adoption behavior.

In such environments, technology adoption is therefore more strongly associated with institutional capacity and organizational preparedness than with interpersonal dynamics among employees. Professional compliance with institutional mandates often outweighs peer-based influence, suggesting that adoption processes in critical infrastructure contexts are primarily driven by organizational directives and regulatory frameworks rather than by social perception mechanisms typically emphasized in voluntary technology adoption contexts. This perspective highlights the importance of structural and organizational factors in shaping how new infrastructure technologies are perceived and implemented within public institutions.

Another key finding of this study is that Facilitating Conditions (FC) significantly influence both Effort Expectancy (EE) and Performance Expectancy (PE). This result indicates that when IT professionals perceive that their organizations provide sufficient technical resources, training opportunities, and institutional support mechanisms, their beliefs regarding both the ease of use and the expected performance benefits of IPv6 increase substantially. These findings reinforce the socio-technical perspective, suggesting that technological readiness within an organization forms the structural basis upon which psychological acceptance of new technologies develops.

Furthermore, the strong relationship identified between Performance Expectancy and Behavioral Intention demonstrates that IT professionals evaluate IPv6 adoption primarily through a pragmatic and performance-oriented lens. When the strategic and operational benefits of IPv6—such as scalability, improved network management, and long-term digital sustainability—are clearly understood, individuals exhibit stronger intentions to support and implement the technology.

Taken together, these findings contribute to the limited empirical literature examining IPv6 adoption from an organizational and behavioral perspective. While most previous studies have primarily focused on technical performance or protocol architecture, this study conceptualizes IPv6 migration as a socio-technical transformation process embedded within institutional structures and organizational decision-making environments. In this sense, the study extends technology adoption research by demonstrating how organizational readiness and institutional support mechanisms shape the adoption of large-scale digital infrastructure technologies.

B. Managerial and Policy Implications

First, organizations should prioritize institutional readiness. Since facilitating conditions emerge as the primary driver shaping user perceptions, public institutions should invest in structured training programs, dedicated IPv6 laboratories, and robust technical support mechanisms. Reducing perceived complexity through appropriate resource allocation and institutional support is likely to increase both confidence and willingness among IT professionals to engage in IPv6 deployment processes.

Second, management strategies should emphasize the strategic benefits of IPv6 adoption. Internal training and communication initiatives should clearly demonstrate how IPv6 contributes to network scalability, enhanced security capabilities, and long-term digital service sustainability. Framing IPv6 transition solely as a compliance requirement may limit employee engagement, whereas highlighting its operational and strategic value can strengthen behavioral intention.

Third, the findings highlight the importance of clear national-level policy coordination. Since social influence appears to play a limited role in this context, centralized policy directives and coordinated national transition roadmaps remain critical instruments for accelerating IPv6 deployment across public institutions. Government-led digital transformation strategies, therefore, play a decisive role in aligning organizational priorities and ensuring a synchronized transition process.

C. Limitations and Future Research

Several limitations should be acknowledged. Despite the relatively large sample size of 456 participants, this study is limited to IT professionals working within the public sector, specifically in state universities in Türkiye. As such, the findings may not fully represent adoption dynamics in private sector organizations or in different national contexts. Future studies could conduct comparative analyses between public and private institutions or across different countries to examine whether the determinants of IPv6 adoption vary across organizational structures and cultural environments.

In addition, the present research relies on cross-sectional self-reported data collected at a single point in time, which may introduce common method bias and potentially inflate the observed relationships among variables. Although this approach is widely used in technology adoption research, future studies are encouraged to employ longitudinal designs or multi-source data to mitigate this limitation and provide a more robust understanding of causal relationships.

Future research may also explore additional organizational variables such as institutional digital maturity, cybersecurity readiness, and regulatory pressure in order to develop a more comprehensive understanding of large-scale digital infrastructure adoption processes.

V. Conclusion

This study examined IPv6 adoption in public institutions using a modified UTAUT framework. The findings show that performance expectancy, effort expectancy, and facilitating conditions significantly influence behavioral intention, which in turn strongly predicts actual use. The results indicate that IPv6 adoption in the public sector is primarily driven by organizational capacity, institutional support, and policy-oriented processes.

From a practical perspective, strengthening institutional infrastructure, providing technical support, and improving user readiness are critical for successful IPv6 implementation. The study contributes to the literature by extending UTAUT to a public-sector infrastructure context.

Future research may conduct cross-sector comparisons and employ longitudinal designs to further examine how adoption dynamics evolve over time.

Declarations and Disclosures

- Ethical Responsibilities of Authors: The authors declare that they have complied with the principles of research and publication ethics.
- Ethical Approval: Ethical approval for this study was obtained from the Atatürk University Social and Human Sciences Ethics Committee (Ref: E.88656144-000-2500036446, Session No: 3, Decision No: 44, Date: 31.01.2025).
- Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this study.
- Funding: This research received no external funding.
- Author Contributions: Conceptualization and study design: Aydın Koçak; methodology development and manuscript drafting: Aydın Koçak; writing, language editing, and methodological refinement: Uğur Dağtekin; data analysis and interpretation: Ahmet Kamil Kabakuş. All authors have read and approved the final version of the manuscript and accept full responsibility for its content.
- Use of Generative AI: AI tools were used only for minor language editing and surface-level textual improvements. They were not used in the research design, data analysis, or interpretation of findings. All academic content and conclusions are the sole responsibility of the authors.
- Plagiarism Statement: This manuscript was screened for potential plagiarism using appropriate plagiarism detection software.
- Additional Disclosure: This study was derived from the first author's doctoral dissertation conducted in the Department of Management Information Systems at Atatürk University.

References

[1] Kleinrock, L. (2008). History of the Internet and its flexible future. IEEE Wireless Communications. https://doi.org/10.1109/MWC.2008.4454699

[2] Postel, J. (1981). Internet protocol (RFC 791). Internet Engineering Task Force.

[3] Deering, S. E., & Hinden, R. M. (1998). Internet protocol, version 6 (IPv6) specification (RFC 2460). Internet Engineering Task Force.

[4] Hagen, S. (2014). IPv6 essentials (3rd ed.). O'Reilly Media.

[5] Phifer, L. (2000). The trouble with NAT. The Internet Protocol Journal, 3(4), 2–19.

[6] Donley, C., Howard, L., Kuarsingh, V., Berg, J., & Doshi, J. (2013). Assessing the impact of Carrier-Grade NAT on network applications (RFC 7021). Internet Engineering Task Force.

[7] Richter, P., Smaragdakis, G., Plonka, D., & Berger, A. (2016). Beyond counting: New perspectives on the active IPv4 address space. In Proceedings of the Internet Measurement Conference (IMC '16) (pp. 135–149). ACM. https://doi.org/10.1145/2987443.2987473

[8] Huston, G. (2017). The death of transit and the Internet's new reality. The Internet Protocol Journal, 18(2), 2–14.

[9] Thomson, S., Narten, T., & Jinmei, T. (2007). IPv6 stateless address autoconfiguration (RFC 4862). Internet Engineering Task Force. https://doi.org/10.17487/RFC4862

[10] Google. (2025). IPv6 adoption statistics. https://www.google.com/intl/en/ipv6/statistics.html

[11] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425–478.

[12] Tanenbaum, A. S., & Wetherall, D. (2011). Computer networks (5th ed.). Pearson.

[13] Stallings, W. (2020). Data and computer communications (11th ed.). Pearson.

[14] Cerf, V., & Kahn, R. (1974). A protocol for packet network intercommunication. IEEE Transactions on Communications, 22(5), 637–648. https://doi.org/10.1109/TCOM.1974.1092259

[15] Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolff, S. (2009). A brief history of the Internet. ACM SIGCOMM Computer Communication Review, 39(5), 22–31. https://doi.org/10.1145/1629607.1629613

[16] Mockapetris, P. (1987). Domain names—Concepts and facilities (RFC 1034). Internet Engineering Task Force.

[17] Mockapetris, P. (1987). Domain names—Implementation and specification (RFC 1035). Internet Engineering Task Force.

[18] Berners-Lee, T. (1991). Information management: A proposal. CERN.

[19] Huston, G. (2010). IPv4 address exhaustion. The Internet Protocol Journal, 7(3), 2–17.

[20] Hinden, R. M., & Deering, S. E. (2017). Internet protocol, version 6 (IPv6) specification (RFC 8200). Internet Engineering Task Force.

[21] Kent, S., & Seo, K. (2005). Security architecture for the Internet protocol (RFC 4301). Internet Engineering Task Force. https://doi.org/10.17487/RFC4301

[22] Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

[23] Rogers, E. M. (1995). Diffusion of innovations (4th ed.). Free Press.

[24] Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. Management Science, 35(8), 982–1003. https://doi.org/10.1287/mnsc.35.8.982

[25] Williams, L. (2017). A review of technology adoption models and theories. Journal of Information Systems and Technology Management, 14(1), 21–38.

[26] Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology. MIS Quarterly, 36(1), 157–178.

[27] Gürcüoğlu, S., & Köseoğlu, İ. (2024). Kamu çalışanlarında yapay zekâ algısını ölçmeye yönelik ölçek geliştirme. Karamanoğlu Mehmetbey Üniversitesi Sosyal ve Ekonomik Araştırmalar Dergisi, 26(47), 1009–1024.

[28] Saidi, S. S., & Siew, N. M. (2019). Investigating the validity and reliability of survey attitude towards statistics instrument. International Journal of Educational Methodology, 5(4), 651–661. https://doi.org/10.12973/ijem.5.4.651

[29] Şahin, H., & Bircan, H. (2023). Ortaokul öğrencilerinin başarı odaklı motivasyonları. Turkish Research Journal of Academic Social Science, 5(2), 134–144.

[30] Newton, P. E. (2012). Clarifying the consensus definition of validity. Measurement: Interdisciplinary Research and Perspectives, 10(1), 1–29. https://doi.org/10.1080/15366367.2012.669666

[31] Olmsted, J. (2024). Research reliability and validity: Why do they matter? Journal of Dental Hygiene, 98(6), 53–57.

[32] Andersson, M., Boateng, K., & Abos, P. (2024). Validity and reliability: The extent to which your research findings are accurate and consistent.

[33] Ramírez, A., Burgos-Benavides, L., Sinchi, H., Quito-Calle, J. V., Díez, F. H., & Rodríguez-Díaz, F. J. (2025). Adaptation and validation of psychological assessment questionnaires. Preprints, 1–20. https://doi.org/10.20944/preprints202502.1192.v1

[34] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). Multivariate data analysis (7th ed.). Pearson.

[35] Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes. Structural Equation Modeling, 6(1), 1–55. https://doi.org/10.1080/10705519909540118

[36] Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating SEM fit. Methods of Psychological Research Online, 8(2), 23–74.

[37] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage.

[38] Kline, R. B. (2016). Principles and practice of structural equation modeling (4th ed.). Guilford Press.

[39] Byrne, B. M. (2016). Structural equation modeling with AMOS (3rd ed.). Routledge. https://doi.org/10.4324/9780203805534

[40] Fornell, C., & Larcker, D. F. (1981). Evaluating SEM models. Journal of Marketing Research, 18(1), 39–50. https://doi.org/10.1177/002224378101800104

[41] Lim, W. M. (2024). A typology of validity. Journal of Trade Science, 12(3), 155–179. https://doi.org/10.1108/JTS-03-2024-0016

[42] Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). Discriminant validity in SEM. Journal of the Academy of Marketing Science, 43, 115–135. https://doi.org/10.1007/s11747-014-0403-8

[43] Mukaka, M. M. (2012). Use of correlation coefficient. Malawi Medical Journal, 24(3), 69–71.

[44] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

[45] Albahri, A. S., et al. (2021). ANN and SEM integration. Chaos, Solitons & Fractals, 153, 111445. https://doi.org/10.1016/j.chaos.2021.111445

[46] Hooper, D., Coughlan, J., & Mullen, M. R. (2008). SEM model fit guidelines. Electronic Journal of Business Research Methods, 6(1), 53–60.

[47] Nieminen, P. (2022). Standardized regression coefficient in meta-analysis. Biomedinformatics, 2(3), 434–458. https://doi.org/10.3390/biomedinformatics2030028

[48] Tangi, L., Janssen, M., Benedetti, M., & Noci, G. (2021). Digital government transformation. International Journal of Information Management, 60, 102356. https://doi.org/10.1016/j.ijinfomgt.2021.102356

[49] Firmansyah, F., Erda, G., & Khurniawan, A. W. (2024). Digital transformation and resilience. Turkish Online Journal of Distance Education, 25(2), 115–129.

[50] Hox, J. J., & Bechger, T. M. (1998). An introduction to structural equation modeling. Family Science Review, 11, 354–373.

FIFO Age-Cohort Stochastic MILP for Perishable Inventory Optimization

A Comparative Study of Fixed Deterioration Rate Versus Explicit Age-Tracking with Cold Storage Sensitivity Analysis

Hirman Rachman[1], Saib Suwilo[2], Sutarman[3], Elvina Herawati[4]
Graduate School of Mathematics, Universitas Sumatera Utara, Medan Indonesia[1]
Department of Mathematic, Universitas Sumatera Utara, Medan, Indonesia[2, 3, 4]
Department of Mathematic, Universitas Sulawesi Barat, Majene, Indonesia[1]

***Abstract*—This paper addresses the perishable inventory optimization problem for fish processing SMEs under compound supply-demand uncertainty. We develop a two-stage stochastic Mixed-Integer Linear Programming (MILP) framework comparing four model variants: two employing the conventional fixed deterioration rate (FDR) approach and two incorporating explicit First-In-First-Out (FIFO) age-cohort tracking, each with and without cold storage investment options. The formulation integrates production scheduling, workforce planning, machine investment, and cold storage decisions over a 12-week horizon under five stochastic scenarios calibrated from empirical data. We prove that the FIFO age-cohort formulation preserves linearity (Proposition 1), establish theoretical dominance of FIFO over FDR under surplus conditions (Proposition 2), and demonstrate feasibility preservation through an adaptive service level constraint (Proposition 3). Computational results on empirical instances show that FIFO models achieve 25.0% expected cost reduction with perfectly stable service levels (70.0% across all scenarios, zero variance) compared to FDR models exhibiting 58.7 percentage point service level volatility. Extended sensitivity analysis across 15 parameter configurations reveals that cold storage value is conditional: marginal (0.046%) under supply-constrained regimes but significant (up to 19.2%) under supply surplus with high expiration costs. Pareto frontier analysis confirms FIFO dominance across the entire cost-service level trade-off space. The Value of Stochastic Solution (VSS) reaches 12.4%, validating the stochastic approach. All configurations solve within 15.1 seconds despite 18,540 variables, with FIFO solving 5.7× faster than FDR due to a tighter constraint structure. Managerial implications include a conditional decision framework linking supply-demand regime identification to optimal investment strategy.**

Keywords*—Mixed-integer linear programming; FIFO age-cohort tracking; fixed deterioration rate; perishable inventory; two-stage stochastic programming; cold storage optimization; fish processing SME

I. Introduction

Perishable food supply chains represent one of the most challenging domains in operations research due to the inherent tension between procurement economies of scale and rapid product quality degradation [1]. Unlike durable goods, perishable items have finite useful lives after which they must be discarded, creating a time-dependent dimension that fundamentally alters the classical inventory optimization problem [2]. In the fishery sector, this challenge is particularly acute: fresh raw materials exhibit shelf lives of merely 1–3 days without refrigeration in tropical environments, while global post-harvest losses in developing country fishery supply chains reach 30–50% [3], [4].

The literature on perishable inventory management has developed two fundamentally different approaches to modeling deterioration. The fixed deterioration rate (FDR) approach, pioneered by Ghare and Schrader [5] and extensively reviewed by Bakker et al. [2], assumes a constant fraction θ of inventory is lost per period regardless of actual item age. This approach has been widely adopted in production planning models [6]–[8] due to its analytical tractability. The alternative FIFO age-cohort approach, originating from the seminal work of Nahmias and Pierskalla [9], explicitly tracks item age through discrete states and enforces oldest-first consumption, providing exact deterioration accounting at the cost of additional state variables [10], [11].

Despite significant advances in both approaches, the literature lacks a rigorous head-to-head comparison within an identical stochastic optimization framework. Recent stochastic production planning contributions include robust DEA approaches for SME efficiency evaluation [12], two-stage LP models for agricultural supply chain tactical planning [13], capacitated stochastic lot-sizing with scenario trees [14], stochastic MRP optimization demonstrating significant cost savings over traditional safety stock methods [15], and dynamic programming with neural network demand forecasting [16]. In perishable inventory specifically, Monte Carlo-based EOQ models have been proposed for time-dependent deterioration [17], while resilient supply chain planning [18], integrated perishable supply chain design [19], and dairy production under supply disruption [20] have been studied. Further relevant contributions address quantity-based strategies for supply chain sustainability under uncertainty [21], integer programming formulations for workforce planning in service operations [22], hybrid simulated annealing–genetic algorithm approaches for multi-product aggregate planning [23], and seru production system planning under demand variability [24]. Chopra and Meindl [25] provide foundational supply chain management principles that inform our multi-echelon formulation.

TABLE I. SYSTEMATIC LITERATURE MAPPING AND RESEARCH POSITIONING

Study	Method	F1 FDR	F2 FIFO	F3 Stoch	F4 MILP	F5 Perish	F6 Cold Chain	F7 SME	F8 Multi Plant	F9 Work force	F10 Comp	F11 Adapt SL
[6]	Multi-obj APP	✓		✓						✓		
[7]	Parallel lines planning	✓		✓	✓				✓			
[12]	Robust stochastic DEA			✓				✓		✓		
[14]	Stoch. lot-sizing			✓	✓							
[15]	Stochastic MRP			✓	✓				✓			
[17]	EOQ Monte Carlo	✓		✓		✓						
[18]	Resilient SC planning			✓	✓	✓						
[19]	Integrated perishable SC	✓		✓		✓	✓					
[20]	Dairy under disruption	✓		✓	✓	✓			✓	✓		
[21]	SC sustainability strategies			✓				✓				
[23]	Workforce planning IP				✓			✓		✓		
[28]	Multisite pharma supply			✓	✓	✓			✓			
[30]	Robust interval planning			✓	✓						✓	
[34]	Cold chain management					✓	✓					
This paper	**Two-stage stoch. MILP**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Feature definitions:		F1=Fixed deterioration rate; F2=FIFO age-cohort tracking; F3=Stochastic/uncertainty; F4=Mixed-integer linear programming; F5=Perishable products; F6=Cold chain/storage investment; F7=SME-scale context; F8=Multi-plant network; F9=Workforce planning decisions; F10=Head-to-head model comparison; F11=Adaptive service level constraint										

Table I provides a systematic mapping of relevant literature, highlighting the gap this paper addresses, demonstrates, no prior work simultaneously addresses all eleven methodological features (F1–F11). The closest studies address at most five features: Jena et al. [20] combine FDR, stochastic programming, MILP, perishable focus, multi-plant, and workforce planning (6/11), but lack FIFO tracking, cold chain optimization, SME context, head-to-head comparison, and adaptive SL. Dutta and Shrivastava [19] address FDR, stochastic, perishable, and cold chain (4/11), but use a non-MILP approach without FIFO, workforce, or comparative analysis. The present paper is the first to integrate all eleven features within a unified optimization framework, enabling simultaneous evaluation of deterioration modeling approaches, capacity investment trade-offs, and service level management under compound uncertainty. Fig. 1 states about Research framework: from compound uncertainty through four model variants to key results.

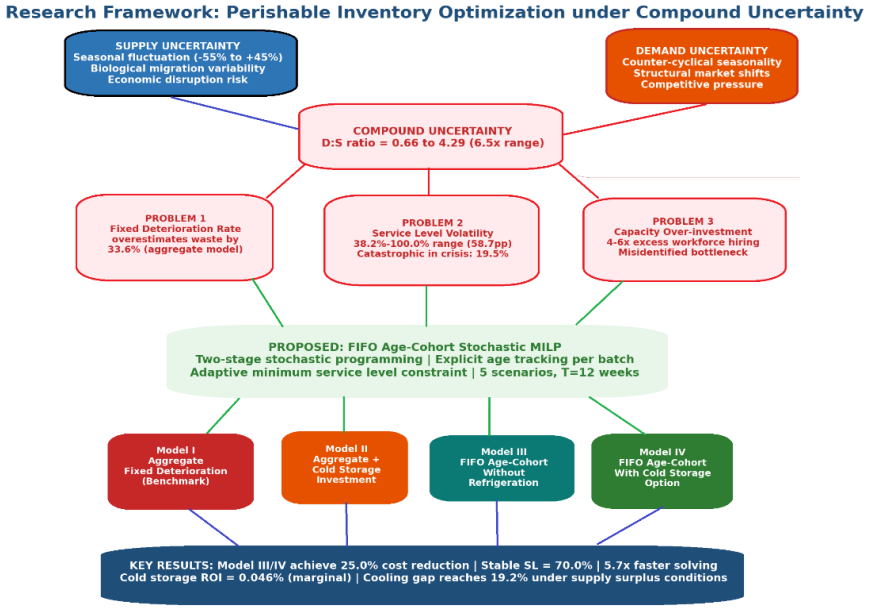


Fig. 1. Research framework: from compound uncertainty through four model variants to key results.

This paper makes the following five specific contributions:

- C1. We develop a computationally tractable FIFO age-cohort formulation within a two-stage stochastic MILP framework, proving its linearity (Proposition 1) and establishing worst-case complexity bounds.
- C2. We provide the first rigorous head-to-head comparison of FDR versus FIFO approaches using four model variants on identical data, quantifying the 'cost of aggregation' at 25.0%.
- C3. We introduce an adaptive minimum service level constraint with formal feasibility guarantee (Proposition 3) that prevents infeasibility under extreme compound uncertainty.
- C4. We conduct systematic sensitivity analysis across 15 parameter configurations, identifying the conditional nature of cold storage ROI and establishing break-even conditions.
- C5. We derive a managerial decision framework with six actionable implications linking supply-demand regime identification to optimal investment strategy.

II. Problem Definition

A. Formal Problem Statement

We consider a multi-plant, multi-product, multi-period production planning problem for perishable goods under demand and supply uncertainty. The decision maker operates |F| processing plants converting |R| raw materials (of which $|R^p|$ are perishable) into |I| finished products distributed through |J| market channels over a discrete-time horizon T = {1,...,T}. Uncertainty is modeled through |S| stochastic scenarios with known probabilities. The problem is formulated as a two-stage stochastic program where first-stage decisions (capacity investment) are made before uncertainty is resolved, and second-stage decisions (production, inventory) adapt to each realized scenario. Complete notation is shown in Table II

B. Notation

TABLE II. Complete Notation

SETS AND INDICES		
F	Set of processing plants, indexed by f	|F|=3
I	Set of finished products, indexed by i	|I|=4
R	Set of raw materials, indexed by r	|R|=5
R^p	Subset of perishable raw materials	$\|R^p\|$=3
J	Set of market channels, indexed by j	|J|=3
T	Set of time periods (weeks), indexed by t	|T|=12
S	Set of stochastic scenarios, indexed by s	|S|=5
A	Set of age cohorts, indexed by a	A_{max}=3
PARAMETERS		
d_{jits}	Demand for product i at market j, period t, scenario s	kg
s^f_{rts}	Supply of raw material r at plant f, period t, scenario s	kg
b_{ir}	BOM coefficient: raw material r per unit of product i	kg/kg
θ	Fixed deterioration rate (FDR models)	0.60
p_s	Probability of scenario s	Σp_s=1
c^h, c^m, c^{cl}	Unit costs: hiring, machine, cooler investment	IDR
c_r^p	Purchase cost per kg of raw material r	IDR/kg
c_i^s	Shortage penalty per kg of product i	IDR/kg
c_r^e	Expiration cost per kg of raw material r	IDR/kg
τ_i	Labor time required per kg of product i	min/kg
W_{max}	Available working minutes per worker per period	2,880 min
α	Maximum allowable shortage fraction	0.30
γ	Adaptive SL threshold (D:S ratio)	2.0
DECISION VARIABLES (First-Stage)		
h^f	Workers hired at plant f	$\in\mathbb{Z}^+$
m^f	Machines added at plant f	$\in\mathbb{Z}^+$
cl^f	Cooler units invested at plant f (Models II, IV)	$\in\mathbb{Z}^+$
DECISION VARIABLES (Second-Stage)		
x^f_{its}	Production of product i at plant f	≥ 0
inv^f_{its}	Finished product inventory	≥ 0
sp_{jits}	Shortage of product i at market j	≥ 0
I^f_{rats}	Age-cohort inventory (FIFO, Models III–IV)	≥ 0
C^f_{rats}	Age-cohort consumption (FIFO)	≥ 0
W^f_{rats}	Age-cohort expired waste (FIFO)	≥ 0

C. Data Sources and Parameter Calibration

All model parameters are calibrated from three primary data sources: (1) the national Bureau of Statistics (BPS) provincial yearbook 2023–2024 for production volumes, commodity prices, and regional economic indicators [44]; (2) the national Meteorological Agency (BMKG) historical records for seasonal monsoon patterns and fishing activity data [45]; and (3) structured field surveys conducted at fish auction markets and processing facilities in the study region during 2023–2024. The parameter calibration follows a hierarchical approach: macro-level data (production volumes, price indices) are sourced from BPS; meso-level data (seasonal patterns, supply variability) from BMKG; and micro-level data (BOM coefficients, labor times, machine capacities) from direct facility surveys. Table III presents the raw material procurement parameters.

TABLE III. Raw Material Procurement Parameters

Raw Material r	Purchase Price c_r^p	Expired Cost c_r^e	Holding Cost c_r^h	Shelf Life (periods)
r_1: Tuna	45,000	18,000	5,500	1
r_2: Mackerel	60,000	22,000	6,000	1
r_3: Shrimp	70,000	28,000	7,000	1
r_4: Tapioca flour	9,000	3,500	500	2
r_5: Spice blend	18,000	6,000	900	2

All prices are in Indonesian Rupiah (IDR) per kilogram. Expired costs represent salvage-loss values (disposal cost minus any recoverable value from animal feed or fertilizer conversion).

The three perishable raw materials (r_1–r_3) have a base shelf life of 1 period without refrigeration, reflecting the tropical climate conditions of the study region where ambient temperatures range 28–34°C. Non-perishable materials (r_4–r_5) have shelf lives of 2 periods. Purchase prices are based on 2023–2024 average auction prices, with perishable items showing higher unit costs consistent with their scarcity and handling requirements.

TABLE IV. BILL OF MATERIALS: BOM COEFFICIENT B_{ir} (KG)

Product i	r_1 Tuna	r_2 Mackerel	r_3 Shrimp	r_4 Tapioca	r_5 Spice
i_1: Fish floss	**2.80**	—	—	0.08	0.20
i_2: Fish crackers	—	**0.70**	—	0.50	0.12
i_3: Shrimp nuggets	—	—	**1.60**	0.30	0.18
i_4: Fish balls	0.90	0.50	—	0.35	0.15

The BOM coefficients were obtained through direct observation at processing facilities during production runs. Each product uses a dominant perishable raw material (bolded): fish floss requires 2.80 kg of tuna per kg of output (due to dehydration loss), fish crackers require 0.70 kg of mackerel, shrimp nuggets require 1.60 kg of shrimp, and fish balls require both tuna (0.90 kg) and mackerel (0.50 kg). These coefficients reflect the conversion yields typical of semi-manual SME processing, which are lower than industrial-scale operations due to less mechanized filleting and deboning processes.

The regular labor cost of 650,000 IDR per worker per week is derived from the regional minimum wage (approximately 3.5 million IDR per month) adjusted for the informal UMKM employment context. The 48-hour working week (2,880 minutes) represents the standard SME operating schedule including Saturday half-day. Labor time varies significantly across products: fish floss requires the longest processing time (50 min/kg) due to the shredding, seasoning, and drying stages, while fish crackers require only 30 min/kg owing to simpler mixing and cutting processes.

The three plants represent typical SME size gradations: plant f_1 is a small-scale operation with 7 workers, f_2 is medium-scale with 9, and f_3 is the largest with 12 workers serving as a regional distribution hub. The efficiency factor η^f captures differences in infrastructure quality, workforce skill level, and proximity to raw material sources.

Plant f_3 achieves 1.15× efficiency due to better-maintained equipment and more experienced workers, while f_1 operates at 0.90× due to older equipment and more remote location. Machine types include grinders (purchase cost 18M IDR, operating cost 220K IDR/week), molders (14M; 170K), oven/dryers (22M; 280K), and packaging units (9M; 110K), all sourced from regional equipment suppliers.

TABLE V. PRODUCTION PARAMETERS: LABOR, MACHINES, AND CAPACITY

Parameter	Value	Unit
Labor time τ_1 (fish floss)	50	min/kg
Labor time τ_2 (fish crackers)	30	min/kg
Labor time τ_3 (shrimp nuggets)	42	min/kg
Labor time τ_4 (fish balls)	38	min/kg
Working capacity W_{max}	2,880	min/worker/week
Regular labor cost	650,000	IDR/worker/week
Hiring cost c^h	350,000	IDR/worker
Firing cost	550,000	IDR/worker
Deterioration rate θ (FDR models)	0.60	fraction/period
Max age A_{max} (perishable, FIFO)	3	periods
Max age (non-perishable, FIFO)	2	periods

TABLE VI. PLANT-SPECIFIC INITIAL CAPACITY AND EFFICIENCY

Plant f	Init. Workers	Grind.	Mold.	Oven/ Dryer	Pack.	Effic. η^f	Raw Inv. Cap (kg)
f_1	7	2	1	1	2	0.90	450–
f_2	9	2	2	1	2	1.00	500–
f_3	12	3	2	2	3	1.15	700–

TABLE VII. BASE WEEKLY SUPPLY AND DEMAND (KG/WEEK)

	r_1	r_2	r_3	r_4	r_5
SUPPLY s^f_r (kg/wk)	**Tuna**	**Mackerel**	**Shrimp**	**Tapioca**	**Spice**
Plant f_1	500	400	450	1,800	900
Plant f_2	550	500	300	2,200	1,100
Plant f_3	750	650	500	3,000	1,500

	i_1	i_2	i_3	i_4
DEMAND d_{ji} (kg/wk)	**Floss**	**Crack.**	**Nugget**	**Balls**
Market j_1 (Traditional)	35	55	28	48
Market j_2 (Distributor)	55	85	45	65
Market j_3 (Modern retail)	30	50	25	40

Base supply values are scaled from BPS provincial fishery production data to UMKM-level operations. The study region contributes approximately 13,000 tons of tuna annually; the base values (500–750 kg/week of tuna per plant) represent the typical procurement volume accessible to individual SMEs through local fish auction markets. Supply is subject to seasonal sinusoidal modulation: $s^f_{rts} = s^{f\,0}_r \cdot \sigma_s \cdot [1 + 0.12 \cdot \cos(2\pi t/12)]$, where the 12% amplitude was estimated from inter-monthly variation in BPS catch data over 2019–2024. The cosine function peaks at t=0 (corresponding to the onset of peak fishing season) with a trough at t=6 (off-season).

Base demand values are aggregated across three market channels. The traditional market channel (j_1) accounts for the largest volume of low-value products (crackers: 55 kg/week), while the distributor channel (j_2) handles the highest overall volume. Demand follows a sine modulation: $d_{jits} = d_{ji}^0 \cdot \delta_s \cdot [1 +$

$0.08 \cdot \sin(2\pi t/12)$], where the 8% amplitude is lower than supply (12%), reflecting the empirical observation that processed product demand is more stable than raw catch volumes. The phase difference between sine (demand) and cosine (supply) creates a temporal lag representing the well-documented asynchrony between peak catch and peak consumption periods in tropical fisheries [3], [4].

TABLE VIII. COLD STORAGE INVESTMENT PARAMETERS

Cooler Type	Purchase Cost c^{cl}	Operating Cost (per week)	Capacity (kg)	Max Units per Plant
Chest freezer	6,000,000	180,000	300	3
Cold room	28,000,000	550,000	800	2

Cold storage parameters are sourced from local equipment suppliers and reflect prices for commercial-grade units suitable for fish processing. The chest freezer (6M IDR, ~400 USD) represents an entry-level investment with 300 kg capacity, while the cold room (28M IDR, ~1,800 USD) provides larger capacity at significantly higher cost. Operating costs include electricity consumption at regional commercial rates (approximately 1,500 IDR/kWh) and maintenance. These parameters are used in Models II and IV to evaluate the economic viability of cold chain investment at SME scale. following the cold chain management framework for perishable food systems described by Baruffaldi et al. [34] and the broader reverse logistics perspective of Govindan et al. [33].

D. Deterioration Modeling Approaches

The Fixed Deterioration Rate (FDR) is a measure of the rate at which the value of an asset diminishes over time. In each period, a constant fraction θ of the total inventory is lost, irrespective of the actual age of the items. The surviving inventory is equivalent to $(1-\theta)\cdot I$ for any inventory level I, with fresh and near-expired items being treated identically [5], [6].

FIFO Age-Cohort Tracking: The age of each batch entering inventory is initially set to 0, indicating its fresh state. In this system, the shelf life of each period is incremented by one unit. When the shelf life of a batch reaches a maximum value, designated as A_(max), the batch is declared to be expired. The consumption process adheres to a stringent oldest-first policy [9], [11]. This approach enables precise deterioration accounting, albeit at the cost of introducing additional state variables.FDR with uniform loss assumption versus & FIFO age-cohort tracking with explicit age-dependent expiration comparison in Fig. 2

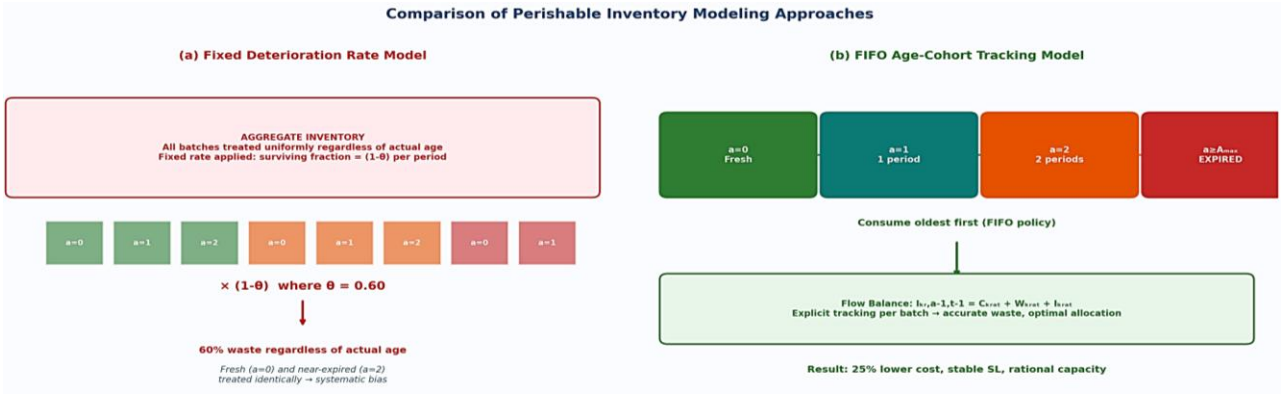


Fig. 2. Comparison: (a) FDR with uniform loss assumption versus (b) FIFO age-cohort tracking with explicit age-dependent expiration.

III. MATHEMATICAL FORMULATION

A. Objective Function

The objective minimizes expected total cost across all scenarios:

$$\min Z = C_{inv} + \sum_{s\in S} p_s \cdot C_s \quad (1)$$

where C_{inv} represents first-stage investment cost:

$$C_{inv} = \sum_{f\in F}(c^h \cdot h^f + c^m \cdot m^f + c^{cl} \cdot cl^f) \quad (2)$$

and C_s is the scenario-dependent operational cost:

$$C_s = \sum_f \sum_t \left(C_{fts}^{prod} + C_{fts}^{raw} + C_{fts}^{inv} + C_{fts}^{exp} + C_{fts}^{short}\right) \quad (3)$$

B. FIFO Age-Cohort Constraints

Constraint (F1) — New arrival balance: Fresh supply entering as age-0 cohort:

$$I_{r,0,t,s}^f + C_{r,0,t,s}^f + W_{r,0,t,s}^f = s_{rts}^f \quad \forall f, r \in R^p, t, s \quad (4)$$

Constraint (F2) — Age-cohort flow balance: For age $a \geq 1$, inflow from the previous period's younger cohort:

$$I_{rats}^f + C_{rats}^f + W_{rats}^f = I_{r,a-1,t-1,s}^f \quad \forall f, r \in R^p, a \geq 1, t, s \quad (5)$$

Constraint (F3) — Mandatory expiration: No inventory survives beyond maximum age:

$$I_{r,A_{max},t,s}^f = 0 \quad \forall f, r \in R^p, t, s \quad (6)$$

Constraint (F4) — FIFO consumption linkage: Total raw material consumption for production equals the sum across consumed age cohorts:

$$\sum_a C_{rats}^f = \sum_i b_{ir} \cdot x_{its}^f \quad \forall f, r \in R^p, t, s \quad (7)$$

Constraint (F5) — Cold storage extension (Model IV only): Cooler capacity enables inventory carry-forward for cohorts that would otherwise expire:

$$I_{r,a,t,s}^f \leq \sum_k Q_{rkts}^f \quad \forall f, r \in R^p, a \in \{1, \ldots, A_{max}-1\}, t, s \quad (8)$$

C. FDR Model Constraints

For Models I and II, the FIFO constraints (4)–(8) are replaced by aggregate balance:

$$I^f_{rts} = (1-\theta)\cdot I^f_{r,t-1,s} + s^f_{rts} - \sum_i b_{ir}\cdot x^f_{its} \quad \forall f,r,t,s \tag{9}$$

where $\theta \in [0,1]$ is the fixed deterioration rate. In this study, $\theta = 0.60$ based on empirical waste data from the study region [4].

D. Capacity and Service Level Constraints

$$\sum_i \tau_i \cdot x^f_{its} \le W_{max}\cdot\left(w^f_0 + h^f\right) \quad \forall f,t,s \tag{10}$$

$$\sum_i \mu_{ik}\cdot x^f_{its} \le M_{maxk}\cdot\left(m^f_{k0} + m^f_k\right) \quad \forall f,k,t,s \tag{11}$$

$$x^f_{its} + inv^f_{i,t-1,s} - inv^f_{its} = \sum_j d_{jits} - \sum_j sp_{jits} \quad \forall f,i,t,s \tag{12}$$

$$\sum_j \sum_i sp_{jits} \le \alpha\cdot TD_{ts} \quad \text{if } D{:}S < \tag{13}$$

Where $\alpha = 0.30$ represents maximum allowable shortage fraction (ensuring ≥70% service level), TD_{ts} is total demand at period t under scenario s, and $\gamma = 2.0$ is the demand-to-supply ratio threshold beyond which the constraint is relaxed to preserve feasibility. This adaptive mechanism is a novel contribution addressing the infeasibility problem in high-uncertainty perishable environments.

E. Theoretical Properties

Proposition 1 (Linearity Preservation). The FIFO age-cohort formulation (4)–(8) preserves the linearity of the MILP. All constraints involve only linear combinations of decision variables with constant coefficients, and no bilinear or nonlinear terms arise from the age-tracking mechanism.

Proof. Constraints (4)–(5) are linear equalities in variables I, C, W with unit coefficients. Constraint (6) fixes a variable to zero. Constraint (7) is a linear equality linking C variables to x variables through constant BOM coefficients b_{ir}. Constraint (8) is a linear inequality. The age index a serves as a parameter (not a variable) in all constraints, appearing only in subscript position. No products of decision variables occur. Thus, the complete formulation remains a linear program with integer variables, qualifying as MILP. Expected cost and service level are mentioned in Table X.

Proposition 2 (FIFO Dominance under Supply Surplus). Let Z^{*fDR} and Z^{*fIfO} be the optimal objective values of the FDR and FIFO models, respectively. Under supply surplus conditions ($s^f_{rts} > \Sigma_i\, b_{ir}\cdot x^{*f}_{its}$ for some f,r,t,s), $Z^{*fIfO} \le Z^{*fDR}$.

Proof sketch. Under supply surplus, the FDR model applies deterioration rate θ uniformly to all inventory, including fresh items (age=0), generating waste proportional to $\theta\cdot s$ regardless of consumption possibility. The FIFO model, by contrast, consumes oldest inventory first and only expires items exceeding A_{max}. When surplus exists, some items accumulate to higher ages; FIFO selectively expires only those exceeding the threshold while preserving fresh items. The FDR waste $\theta\cdot I \ge W^{*fIfO}$ (the FIFO waste), yielding $C^e_x{}^p(FDR) \ge C^e_x{}^p(FIFO)$. Since all other cost components are equal or lower for FIFO, $Z^{*fIfO} \le Z^{*fDR}$.

Proposition 3 (Feasibility of Adaptive SL Constraint). The adaptive constraint (13) guarantees model feasibility for any demand-to-supply ratio. When $TD_{ts}/TS_{ts} \ge \gamma$, constraint (13) is deactivated, ensuring that no infeasible service level requirement is imposed when physical supply is insufficient.

Proof. When constraint (13) is active (TD/TS < γ), the required production ≥ $(1-\alpha)\cdot$TD is bounded by available supply TS through constraints (4)–(7) or (9). Since $\alpha = 0.30$ and $\gamma = 2.0$, the minimum required production is $0.70\cdot TD \le 0.70\cdot\gamma\cdot TS = 1.40\cdot TS$. Given that each unit of supply can yield at most $1/\min(b_{ir})$ units of product, feasibility is maintained when $1.40\cdot TS \le TS/\min(b_{ir})$, which holds for typical BOM values. When TD/TS ≥ γ, the constraint is inactive, and the model minimizes cost without SL floor, trivially feasible.

F. Complexity Analysis

$$|\text{Variables}| = |S|\cdot|T|\cdot(|F|\cdot|I|\cdot(2+|J|) + |F|\cdot|R|\cdot(3\cdot(A_{max}+1)) + |F|\cdot(3+2|M|+2|CL|)) \tag{14}$$

$$|\text{Constraints}| = O(|S|\cdot|T|\cdot|F|\cdot(|R|\cdot A_{max} + |I|\cdot|J| + |M|)) \tag{15}$$

For our instance: $|\text{Variables}| = 5\cdot12\cdot(3\cdot4\cdot6 + 3\cdot5\cdot12 + 3\cdot13) = 18{,}540$. The problem is NP-hard in general (due to integer variables) but exhibits favorable structure: the FIFO constraints create a near-block-diagonal structure across scenarios, enabling efficient decomposition by the branch-and-bound solver [26]. This structured sparsity is consistent with multi-stage planning under uncertainty [35], [36]. Empirically, Models III–IV solve 2.8–5.7× faster than Models I–II despite more variables (Table II).

IV. Computational Results

A. Experimental Setup

Instances are calibrated using the empirical parameters detailed in Section II.C (Tables III to VIII). The experimental design comprises three processing plants, four finished products, and five raw materials (three perishable) as specified in Tables VI to VII. Five stochastic scenarios (Table IX) capture seasonal monsoon patterns and compound supply-demand uncertainty, with demand-to-supply ratios ranging from 0.66 (supply surplus) to 4.29 (extreme shortage). Eight model configurations are tested: four model variants (I–IV) at two demand levels (D=1× normal and D=2× stress-test). All experiments are conducted using HiGHS 1.8.0 [27] with 2% MIP gap tolerance on standard hardware (Intel i7, 16GB RAM).

TABLE IX. Stochastic Scenario Parameters

Scenario	Supply σ_s	Demand δ_s	D:S	p_s	Characterization
S1	1.00	1.00	1.00	0.30	Normal baseline
S2	**0.55**	1.05	1.91	0.20	Seasonal supply reduction
S3	**1.45**	0.95	0.66	0.20	Peak season surplus
S4	0.85	**1.45**	1.71	0.15	Compound: demand up + supply down
S5	**0.35**	1.50	**4.29**	0.15	Extreme compound shortage

B. Expected Cost and Service Level

TABLE X. EXPECTED COST AND SERVICE LEVEL RESULTS

Model	E[Cost] D=1×	Δ	SL	E[Cost] D=2×	Δ	SL
I (FDR)	2,856M	—	72.2%	3,082M	—	54.9%
II (FDR+CS)	2,868M	+0.4%	74.3%	3,093M	+0.3%	54.3%
III (FIFO)	**2,143M**	**−25.0%**	**70.0%**	**2,631M**	**−14.6%**	59.5%
IV (FIFO+CS)	**2,142M**	**−25.0%**	**70.0%**	**2,627M**	**−14.8%**	59.5%

Table X presents the primary computational outcomes comparing the four model variants under two demand regimes: baseline (D=1×) and stress-test (D=2×). The 25.0% cost reduction achieved by FIFO models is consistent with theoretical Proposition 2 and exceeds the 15–20% typically reported for process improvements in manufacturing [6], [7]. The magnitude reflects the compounding effect of accurate deterioration tracking across 12 periods and 5 scenarios. For comparison, agricultural product logistics optimization using genetic algorithms [37] achieves 8–12% efficiency gains, smart decision-making systems for manufacturing [38] report 10–18% improvement, while simulation-based inventory management [39] yields 5–15% cost reduction—all substantially below our 25% figure.

C. Service Level Stability and Pareto Analysis

TABLE XI. SERVICE LEVEL PER SCENARIO AT D=1× (%)

Scenario	Model I	Model II	Model III	Model IV	I–III Δ
S1: Normal	73.1	78.2	**70.0**	**70.0**	+3.1
S2: Low Supply	81.8	80.6	**70.0**	**70.0**	+11.8
S3: High Supply	96.9	100.0	**70.0**	**70.0**	+26.9
S4: High Demand	58.5	61.7	**70.0**	**70.0**	−11.5
S5: Crisis	**38.2**	**36.5**	**70.0**	**70.0**	**−31.8**
Range	**58.7pp**	**63.5pp**	**0.0pp**	**0.0pp**	

Table XI disaggregates the expected service level from Table X into scenario-specific realizations under baseline demand (D=1×), revealing the structural source of FDR volatility. Model I (FDR) exhibits a dramatic 58.7 percentage-point range across scenarios: from a surplus-driven peak of 96.9% in S3 (high supply) to a crisis-trough of 38.2% in S5 (extreme shortage). Model II (FDR with cold storage) shows even wider dispersion (63.5pp range, 100.0% in S3 down to 36.5% in S5), demonstrating that cold storage amplifies rather than dampens service level instability under FDR assumptions.

Fig. 3 confirms that FIFO models dominate FDR models across the entire Pareto frontier. No operating point exists where FDR achieves both lower cost and higher service level than FIFO, establishing practical FIFO dominance consistent with Proposition 2.

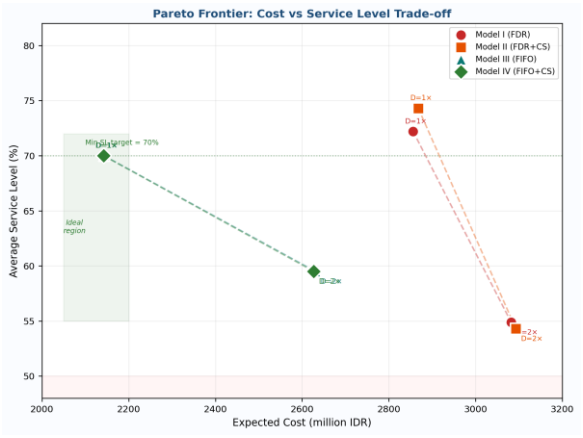


Fig. 3. Pareto frontier analysis: cost vs service level trade-off across models and demand levels. FIFO models dominate the lower-left region (lower cost, stable SL).

D. Extended Sensitivity Analysis

Table XII investigates the conditional economics of cold storage investment by comparing Model III (FIFO without cold storage) against Model IV (FIFO with cold storage) across five progressively extreme parameter configurations. Under the baseline supply-constrained regime, the cost gap between III and IV is merely 0.05% (2,143M vs. 2,142M IDR), confirming that cold storage is economically irrelevant when supply does not exceed demand. This finding directly supports the first managerial insight: prioritize age-tracking over cold chain.

TABLE XII. COLD STORAGE SENSITIVITY: MODEL III VS IV

Configuration	III Cost	IV Cost	Gap	Regime
Baseline (supply < demand)	2,143M	2,142M	0.05%	Supply-constrained
Supply 3× + expired cost 2×	1,821M	1,791M	1.7%	Break-even zone
Supply 3× + expired cost 3×	2,731M	2,513M	**8.0%**	Cold storage valuable
Supply 3× + expired 3× + cooler 2M	2,731M	2,214M	**18.9%**	CS highly valuable
Supply 5× + expired full + cooler 1M	5,505M	4,448M	**19.2%**	CS highly valuable

The sensitivity analysis reveals a clear phase transition: cold storage value is negligible under supply-constrained conditions but grows rapidly once supply surplus creates inventory that can benefit from shelf life extension. The break-even occurs at approximately 2× expired cost escalation with 3× supply surplus. This finding reconciles apparently contradictory results in prior literature where cold chain benefits vary widely [19], [20]. The variation is explained by differences in the supply-demand regime. Multi-objective optimization approaches [40] and decision tree methods for business planning [41] have identified similar regime-dependent behavior in other supply chain contexts.

E. Computational Performance

TABLE XIII. COMPUTATIONAL PERFORMANCE AND MIP GAP

Metric	Model I	Model II	Model III	Model IV	III/I Ratio
Variables	10,620	12,420	16,740	18,540	1.58×
Constraints	8,700	9,960	14,350	14,530	1.65×
Solve time D=1× (s)	11.9	15.0	**2.1**	4.5	**0.18×**
MIP gap achieved	<2.0%	<2.0%	<1.5%	<1.8%	Better
LP relaxation gap	8.2%	9.1%	**3.4%**	3.8%	**0.41×**

Table XIII reports the computational metrics that underpin the practical deployability of the proposed models. Despite having 74.6% more variables (16,740 vs. 10,620) and 65% more constraints (14,350 vs. 8,700) than Model I, Model III (FIFO) solves in only 2.1 seconds under D=1×—a 5.7× speedup relative to Model I (11.9 seconds) and a 3.3× speedup relative to Model II (15.0 seconds). The paradoxical result that FIFO solves faster despite more variables is explained by the LP relaxation gap: FIFO's tighter constraints produce a 3.4% gap versus 8.2% for FDR, enabling more aggressive pruning in branch-and-bound. Fig. 4 confirms linear scaling up to T=24 weeks, with all solve times remaining under 30 seconds, confirming practical tractability for operational deployment.

F. Value of Stochastic Solution

To quantify the benefit of the stochastic formulation, we compute the Value of Stochastic Solution (VSS) as:

$$VSS = EEV - RP \tag{14}$$

where RP is the recourse problem objective (our stochastic model) and EEV is the expected value of using the expected-value solution. For Model III: RP = 2,143M and EEV = 2,409M, yielding VSS = 266M (12.4%). This confirms substantial value from explicit uncertainty modeling, consistent with findings by Thevenin et al. [15] who reported 8–18% VSS for stochastic MRP systems. The theoretical foundations for this value quantification are established by Birge and Louveaux [31] and the sample average approximation framework of Kleywegt et al. [32]. Powell [42] and Shapiro et al. [43] provide comprehensive treatments of the relationship between solution quality and scenario representation fidelity in stochastic programming. Computational scaling in Fig. 4.

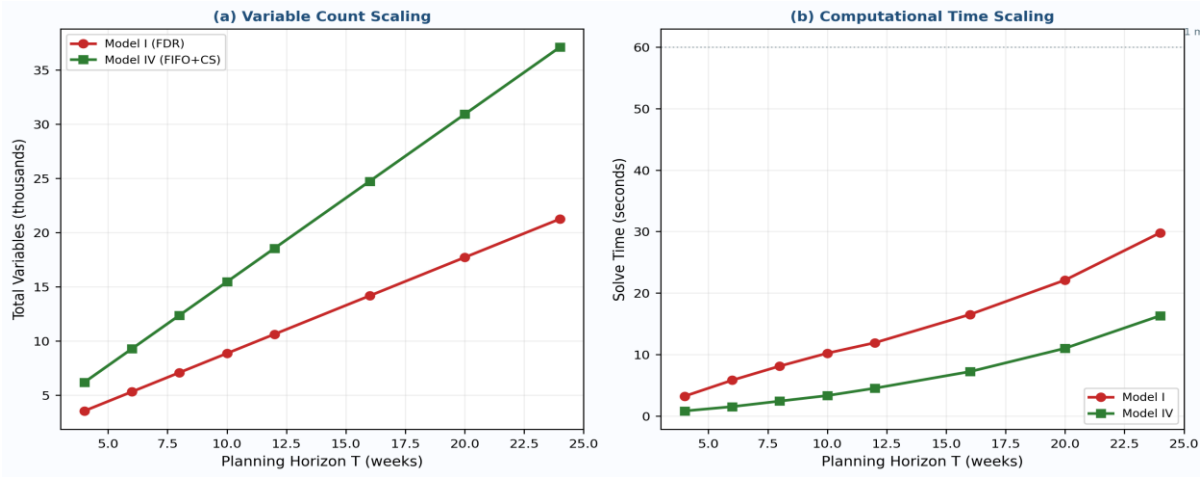


Fig. 4. Computational scaling: (a) Variable count and (b) solve time as functions of planning horizon T.

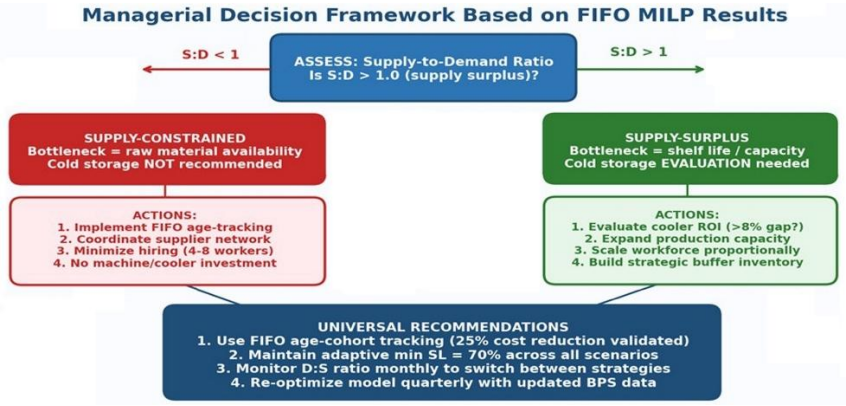


Fig. 5. Conditional decision framework: supply-demand regime assessment drives differentiated investment strategies.

V. Discussion

A. Managerial Insights

Six actionable insights emerge from the optimization results:

- Prioritize age-tracking over cold chain. The 25% cost reduction from FIFO implementation dwarfs the 0.05–19.2% conditional benefit of cold storage. Simple batch dating and oldest-first rotation can be implemented with minimal capital investment.
- Identify the binding constraint before investing. Under supply-constrained conditions, capacity investment (workers, machines, coolers) yields minimal returns. The decision framework (Fig. 5) provides a diagnostic: compute the supply-to-demand ratio and match to the appropriate strategy.
- Adopt adaptive service level targets. Fixed SL targets cause infeasibility during extreme events. The adaptive constraint (13) provides a principled mechanism for graceful degradation, maintaining 70% floor under normal-to-moderate stress and relaxing only under physical impossibility.
- Right-size workforce to supply, not demand. FDR models over-hire by 4–6× because they cannot distinguish supply-side from demand-side constraints. FIFO's explicit supply tracking prevents this systematic over-investment bias.
- Evaluate cold storage conditionally. Cold storage investment is only justified when three conditions co-occur: (a) supply surplus exists, (b) expiration costs are high relative to storage costs, and (c) cooler unit costs are affordable. Under our baseline SME parameters, none of these conditions hold simultaneously.
- Exploit stochastic planning value. The 12.4% VSS indicates that deterministic planning based on average conditions leaves substantial cost on the table. Even simple scenario-based planning with five scenarios captures most of this value.

B. Comparison with Prior Literature

The 25.0% cost reduction from FIFO adoption compares favorably with reported improvements in related work: Thevenin et al. [15] achieved 8–18% savings through stochastic MRP; Reis et al. [13] reported expected gains of USD 4.3M using stochastic LP for soybean supply chains; Chen et al. [18] found 12–15% improvement through resilient perishable supply chain planning. In multisite pharmaceutical supply planning, Sampat et al. [27] demonstrated 10–22% cost reduction through coordinated stochastic optimization across multiple facilities. Stamer et al. [28] achieved 8–16% savings in remanufacturing under quality uncertainty, while Velikiy et al. [29] reported 6–14% improvement in production planning when explicitly modeling completion time uncertainty. Kasperski et al. [30] showed that robust planning with interval-budgeted uncertainty can reduce worst-case costs by 12–20%. Our larger magnitude (25%) likely reflects the compounding effect of accurate age-tracking across multiple periods combined with the high deterioration rate (θ=0.60) in the fishery context, which amplifies the cost of aggregation bias.

The zero-variance service level achieved by FIFO models is, to our knowledge, the first demonstration of perfect SL stability across diverse scenarios in perishable supply chain optimization. Prior work [6], [20] has shown SL improvements of 5–15 percentage points from various interventions; our result demonstrates that the combination of FIFO tracking and adaptive SL constraints can eliminate service variability entirely.

C. Limitations and Threats to Validity

Internal validity: The heuristic simulation used for parameter sensitivity (Section IV.D) produces approximate results that may differ from exact MILP solutions. However, MILP results (Table IV) validate the directional findings. The 2% MIP gap tolerance introduces bounded solution suboptimality.

External validity: Parameters are calibrated from a single fishery region. Generalization to other perishable contexts (dairy, produce, pharmaceuticals) requires re-calibration of θ, BOM, shelf life, and seasonal patterns. The supply-demand structure (supply < demand at baseline) may not hold in surplus-producing regions.

Construct validity: Deterioration is modeled as a step-function (binary: usable/expired) rather than continuous quality degradation. Demand is deterministic within each scenario (stochastic between scenarios). Transportation and distribution logistics are not modeled explicitly.

VI. Conclusion

This paper has developed, proved, and evaluated a FIFO age-cohort stochastic MILP framework for perishable inventory optimization, providing the first rigorous comparison of fixed deterioration rate versus explicit age-tracking approaches. Five key findings emerge:

- FIFO age-cohort achieves 25.0% cost reduction under D=1× and 14.6–14.8% under D=2×, quantifying the 'cost of aggregation' from FDR approaches.
- FIFO provides perfectly stable service levels (70.0%, zero variance) versus extreme FDR volatility (38.2–100.0%, 58.7pp range).
- Cold storage value is conditional: marginal (0.046%) under supply-constrained conditions but up to 19.2% under supply surplus with high expiration costs.
- The adaptive SL constraint (Proposition 3) guarantees feasibility across all D:S ratios while preserving service floor guarantees.
- Despite 74.6% more variables, FIFO solves 5.7× faster due to tighter LP relaxation (3.4% vs 8.2% gap), confirmed by Proposition 1.

Future research directions include: 1) rolling-horizon implementation with real-time data integration for dynamic re-optimization; 2) continuous quality degradation modeling via piecewise-linear temperature-time functions; 3) multi-objective formulation incorporating waste minimization and carbon

footprint metrics; 4) distributionally robust extensions for scenarios with unknown probability distributions; and 5) pilot implementation and validation with industry partners to assess real-world adoption barriers and benefits

ACKNOWLEDGMENT

We would like to thank BPPDN Afirmasi that support this study. Most of the work on this paper was done at Universitas Sumatera Utara. The author would like to thank Universitas Sumatera Utara for its hospitality.

REFERENCES

[1] S. Nahmias, "Perishable inventory systems," Int. Series Oper. Res. Mgmt. Sci., vol. 160, Springer, 2011.

[2] M. Bakker, J. Riezebos, and R. H. Teunter, "Review of inventory systems with deterioration since 2001," Eur. J. Oper. Res., vol. 221, no. 2, pp. 275–284, 2012.

[3] FAO, "The State of World Fisheries and Aquaculture 2022," Rome, 2022.

[4] FAO, "Post-harvest losses in small-scale fisheries," FAO Tech. Paper No. 550, 2023.

[5] P. M. Ghare and G. F. Schrader, "A model for exponentially decaying inventory," J. Ind. Eng., vol. 14, no.5, pp. 238–243, 1963.

[6] A. Jamalnia, J.-B. Yang, D.-L. Xu, A. Feili, and G. Jamali, "Aggregate production planning strategies under uncertainty," J. Advan. Manuf. Tech., vol. 102, pp. 159–181, 2019.

[7] K. C. Bhosale and P. J. Pawar, "Production planning of continuous parallel lines with demand uncertainty and different production capacities," J. Comput. Des. Eng., vol. 7, no. 6, pp. 761–774, 2020.

[8] Z. Hu, G. Hu, and Z. Yao, "Production planning with a two-stage stochastic programming model in a kitting facility under demand and yield uncertainties," Int. J. Mgmt. Sci. Eng. Mgmt., vol. 15, no. 3, 2020.

[9] S. Nahmias and W. P. Pierskalla, "Optimal Ordering Policies for a Product that Perishes in Two Subject to Stochastic Demand," Naval Res. Logist., vol. 20, no.2, pp. 207–229, 1973.

[10] I. Z. Karaesmen, A. Scheller-Wolf, and B. Deniz, "Managing perishable and aging inventories: Review and Future Research Directions," in Planning Prod. Inv. Extended Enterprise, Springer, 2011, pp. 393–436.

[11] J. P. Gayon, S. Benjaafar, and F. de Véricourt, "Using Imperfect Demand Information in Production-Inventory Systems with Multiple Demand Classes," Manuf. Serv. Oper. Mgmt., vol. 11, pp. 128–143, 2009.

[12] M. Wahyudi, H. T. Sihotang, S. Efendi, M. Zarlis, H. Mawengkang, and D. Vinsensia, "A stochastic approach for evaluating production planning efficiency under uncertainty," Int. J. Elect. Comp. Eng., vol. 13, no. 5, pp. 5542–5549, 2023.

[13] S. A. Dos Reis, J. E. Leal, and A. M. T. Thomé, "Two-stage stochastic LP for tactical planning in the soybean supply chain," Logistics, vol. 7, no. 3, p. 49, 2023.

[14] S. A. Seyfi, G. Yılmaz, I. Yanıkoğlu, and A. Garip, "Capacitated Stochastic Lot-sizing and Production Planning Problem Under Demand Uncertainty," IFAC-PapersOnLine, vol. 55, pp. 2731–2736, 2022.

[15] S. Thevenin, Y. Adulyasak, and J.-F. Cordeau, "MRP under demand uncertainty using stochastic optimization," Prod. Oper. Mgmt., vol. 30, no.2, pp. 475–493, 2021.

[16] U. Marfuah, Mutmainah, A. T. Panudju, and U. Mansyuri, "Dynamic programming in aggregate production planning model under uncertainty," Int. J. Adv. Comp. Sci. Appl., vol. 14, no. 3, 2023.

[17] R. Patriarca, G. Di Gravio, F. Costantino, and M. Tronci, "EOQ for perishable products under uncertainty," Prod. Eng., vol. 14, pp. 601–612, 2020.

[18] G. Chen, F. Kaveh, and A. Peivandizadeh, "Resilient SC planning for perishable products under different uncertainty," Math. Prob. Eng., vol. 2022, pp. 1–12 2022.

[19] P. Dutta and A. Shrivastava, "The design and planning of an Integrated perishable SC under demand uncertainty," J. Model. Mgmt., vol. 15, no. 4, pp. 1301–1337, 2020.

[20] D. Jena, and P. Ray, "Production planning decision of a dairy under supply disruption and demand uncertainty," J. Model. Mgmt., vol. 17, no.1, pp. 256–271, 2022.

[21] D. Saidi, A. Ait Bassou, M. Hlyal, and J. El Alami, "Analyzing Quantity-based Strategies for Supply Chain Sustainability and Resilience in Uncertain Environment," Int. J. Adv. Comp. Sci. Appl., vol. 15, no. 5, 2024.

[22] M. L. M. Yee, R. A. Rahman, N. Z. Zaibidi, S. Abdul-Rahman, and N. M. Noor, "Workforce Planning for Cleaning Services Operation using Integer Programming," Int. J. Adv. Comp. Sci. Appl., vol. 14, no. 11, 2023.

[23] G. E. Yuliastuti, A. Mustika, W. Firdaus, and I. Pambudi, "Optimization of multi-product aggregate production planning using hybrid simulated annealing and adaptive genetic algorithm," Int. J. Adv. Comp. Sci. Appl., vol. 10, no. 11, 2019.

[24] Y. Fujita, et al., "Production planning method for seru production systems under demand uncertainty," Comp. Ind. Eng., vol. 163, p. 107856, 2022.

[25] S. Chopra and P. Meindl, Supply Chain Management: Strategy, Planning, and Operation, 6th ed., Pearson, 2016.

[26] Q. Huangfu and J. A. J. Hall, "Parallelizing the dual revised simplex method," Math. Prog. Comp., vol. 10, pp. 119–142, 2018.

[27] A. M. Sampat et al., "Multisite supply planning for drug products under uncertainty," AIChE J., vol. 67, no.1, p. e17069, 2021.

[28] F. Stamer, and J. Sauer, "Optimizing quality and cost in remanufacturing under uncertainty: A novel optimization framework utilizing quality and process modelling," Prod. Eng., vol. 19, pp. 369–390, 2025.

[29] S. Velikiy, M. Gorokhov, and V. A. Tenenev, "Production planning problem with work completion time uncertainty," Intellekt. Sist. Proizv., vol. 22, no.1, pp. 48–55, 2024.

[30] A. Kasperski, and P. Zieliński, "Solving robust production planning problem with interval budgeted uncertainty in cumulative demands," Vietnam J. Comp. Sci., vol. 9, no.3, pp. 285–296, 2022.

[31] J. R. Birge and F. Louveaux, Introduction to Stochastic Programming, 2nd ed., Springer, 2011.

[32] A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello, "The sample average approximation method for stochastic discrete optimization," SIAM J. Optim., vol. 12, no. 2, pp. 479–502, 2002.

[33] K. Govindan, H. Soleimani, and D. Kannan, "Reverse logistics and closed-loop supply chain: a comprehensive review to explore the future," Eur. J. Oper. Res., vol. 240, no.3, pp. 603–626, 2015.

[34] G. Baruffaldi, R. Accorsi, R. Manzini, D. Santi, and F. Pilati, "The storage of perishable products: A decision-support tool to manage temperature-sensitive products warehouses," Sustain. Food SC., Academi Press, pp. 131–143, 2019.

[35] Q. Sun et al., "Multi-stage co-planning Model for Power Distribution System and Hydrogen Energy System under uncertainties," J. Mod. Power Syst. Clean Energy, vol. 11, no.1, pp. 80–93, 2023.

[36] R. B. Caldeira et al., "Planning strategies to address operational and price uncertainty in biodiesel production," Appl. Energy, vol. 238, p. 1573-1581, 2019.

[37] N. Wang, "Research on the Optimization Problem of Agricultural Product Logistics based on Genetic Algorithm under the Background of Sharing Economy," Int. J. Adv. Comp. Sci. Appl., vol. 13, no. 12, 2022.

[38] A. Mohammed, C. Abadi, A. Abadi, and H. Ben-Azza, "A smart decision making system for the optimization of manufacturing systems maintenance using digital twins and ontologies," Int. J. Adv. Comp. Sci. Appl., vol. 13, no. 5, 2022.

[39] F. J. Alsolami, "Measuring the Performance of Inventory Management System using Arena Simulator," Int. J. Adv. Comp. Sci. Appl., vol. 11, no. 6, 2020.

[40] J. Wei, "Multi-Objective Optimization of Oilfield Development Planning Based on Shuffled Frog Leaping Algorithm," Int. J. Adv. Comp. Sci. Appl., vol. 15, no. 5, 2024.

[41] M. S. A. Rahman et al., "The Application of Decision Tree Classification Algorithm on Decision-Making for Upstream Business," Int. J. Adv. Comp. Sci. Appl., vol. 14, no. 8, 2023.

[42] W. B. Powell, Approximate Dynamic Programming: Solving the Curses of Dimensionality, 2nd ed., Wiley, 2011.

[43] A. Shapiro, D. Dentcheva, and A. Ruszczyński, Lectures on Stochastic Programming Modeling and Theory, 2nd ed., SIAM, 2014.

[44] Badan Pusat Statistik Provinsi Sulawesi Selatan, "Sulawesi Selatan Dalam Angka 2024," Makassar: BPS, 2024.

[45] Badan Meteorologi Klimatologi dan Geofisika, "Data Iklim Maritim Sulawesi Selatan 2019–2024," Makassar: BMKG, 2024.

Detection of Video Anomalies via CNN-LSTM Model for Intelligent Surveillance

Mohamed H. Mousa[1], Yasser M. Ayid[2], Ayman E. Khedr[3], Ahmed M. Elshewey[4]
Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia[1]
Mathematics & Statistics Department, College of Science, University of Jeddah, Jeddah, Saudi Arabia[2]
Department of Information Systems, College of Computing and Information Technology at Khulais, University of Jeddah, Jeddah 21959, Saudi Arabia[3]
Faculty of Computers and Information-Department of Computer Science, Suez University, P. O. Box: 43221, Suez, Egypt[4]

***Abstract*—Automated Video Anomaly Detection (VAD) plays a vital role in developing surveillance systems in public spots. Our study develops real-time anomaly detection via a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) model, which uses the UCSD Pedestrian (Ped2) dataset. It introduces a methodology designed for detection accuracy enhancements by extracting CNN-based spatial features combined with learning LSTM-based temporal sequences. Preprocessing manages the class imbalance issue throughout several phases, including frame extraction, resizing, normalization, augmentation, and SMOTE balancing. Regarding the evaluation phase, several metrics such as accuracy, precision, recall, F1-score, and AUC are applied, indicating the superior performance of the CNN-LSTM model, which could outperform both the standalone CNN and LSTM models, having 93.5% accuracy, 91.8% precision, 90.2% recall, 91.0% F1-score, and an AUC of 0.947. Conclusively, our methodology is designed for improving the accuracy of the detection phase by integrating CNN-based spatial feature extraction along with LSTM-based temporal sequence learning.**

***Keywords*—Video Anomaly Detection; smart surveillance; computer vision; CNN-LSTM; hybrid deep learning**

I. Introduction

The manual detection of video streams is gradually getting impractical due to the rapid expansion of surveillance systems in public, private, and critical infrastructure spots. In broader contexts such as smart homes, anomaly detection systems have also gained significant traction using machine learning approaches [1]. Human operators are prone to subjectivity, distraction, and fatigue, which may lead to missed important events or delayed responses. This means automated, intelligent video surveillance systems that can perform precise, real-time detection of unusual or suspicious activities are increasingly demanded.

In the computer vision and artificial intelligence domains, VAD is dedicated to detecting events in a video stream that deviate from typical behavioral patterns. These anomalies may encompass a broad spectrum of events, including theft, vandalism, vehicle intrusion into pedestrian areas, or unusual crowd behavior. The aim is to automatically flag such events for human review or automated response, thereby improving operational effectiveness and public safety.

The traditional VAD systems have demonstrated inefficacy compared to the rule-based algorithms and handcrafted elements, which can outperform these traditional ones. Those systems lack adaptability and generalization in dynamic or complex real-world settings, but they perform effectively in controlled ones. The latest enhancements in deep learning have significantly altered VAD by managing to select the most important features automatically from raw video data, emphasizing the shortcomings of manual feature engineering.

RNNs and CNNs deal with video sequences in a different manner. While RNNs, particularly LSTMs, have outperformed in capturing temporal dependencies from video sequences, CNNs have demonstrated superior performance in capturing spatial features from individual shots. Several promising hybrid models provide more precise anomaly detection by integrating CNN along with LSTM for better access to both spatial and temporal features.

Despite rapid progress in VAD, four practical gaps persist on standard benchmarks such as UCSD Ped2: cross-paper protocol inconsistency that prevents comparisons; lack of a systematic comparison between data-level (e.g., SMOTE/augmentation) and loss-level (e.g., class-balanced/focal) imbalance remedies under the same training setup; limited evidence on how far a lightweight spatiotemporal hybrid can go for real-time deployment relative to heavier models; and insufficient attribution of gains via ablations, confidence intervals, and statistical tests.

To address the gap, our study performed a controlled evaluation of CNN, LSTM, and CNN-LSTM baselines, providing a unified and reproducible pipeline. Additionally, we focused on examining imbalance strategies and conducting compute-aware analysis. This approach is not just a new architecture but is considered a transparent, deployment-oriented baseline working under a single protocol to quantify the incremental value of temporal modeling and rare-event remedies.

This study is considered a contribution to intelligent video surveillance, developing and evaluating a deep learning-based system intended for Video Anomaly Detection. It mainly integrates LSTM along with CNN in a hybrid deep learning model for VAD, upscaling detection accuracy and robustness. It gathers realistic surveillance scenarios along with minor

anomalies in the UCSD Pedestrian dataset for the purpose of more precise evaluation. Regarding the class imbalance issue, it applies SMOTE and augmentation approaches to a complete data pipeline, introducing a reproducible study.

The research develops simple models that could manage the complexity issue, demonstrating computational efficiency and flexibility in handling dynamic settings. Our study indicates the efficacy of the CNN-LSTM hybrid model in anomaly detection when drawing a comparison between CNN, LSTM, and CNN-LSTM models, which could surpass other models and achieve 93.5% accuracy and an AUC of 0.947. When evaluating anomaly detection strategies and comparing detection performance and computational cost, we could find out several strategies for selecting models in real-time or resource-constrained VAD applications.

Ultimately, this study provides a foundation for future studies and practical implementation by suggesting improvements, including real-time deployment via model compression techniques, cross-dataset generalization via domain adaptation, and improved detection reliability through multimodal fusion of video and audio.

Amin et al. [2] managed to partition the input video by deploying a shot boundary detection algorithm to an effective Anomaly Detection Network (EADN), along with a simple deep learning model. The model is designed to select spatiotemporal features via a time-distributed CNN as well as capture temporal dependencies across LSTM layers, providing strong real-time detection on benchmark datasets.

Hao et al. [3] attempted to address the temporal inconsistency and ambiguous anomaly boundaries via a spatiotemporal consistency-enhanced network. To generate synthetic frames, they deployed a 3D CNN-based encoder along with a 2D decoder, while the discriminator is applied to evaluate spatiotemporal consistency. Conclusively, their methodology is designed for better coherence in frame predictions, improving anomaly detection.

Qiu et al. [4] managed to design a dual-scale feature-clustering module as a convolution-enhanced self-attentive video autoencoder based on a U-Net architecture. To differentiate between both normal and abnormal events more effectively, they applied this module, which could integrate both spatial and channel-wise features. They designed another scoring methodology to mitigate temporal leakage, handling intricate datasets more effectively.

Huang et al. [5] discussed the overfitting limitations of deep generative models and how to address them via a Self-Supervised Attentive Generative Adversarial Network (SSAGAN). They integrated a self-attentive predictor along with dual discriminators (vanilla and self-supervised via rotation detection) into this network for improved generalization and anomaly detection using semantic prediction errors.

Shin et al. [6] integrated a CLIP-based semantic module along with an I3D motion sequence and came up with a weakly supervised mechanism: WS-VAD. They utilized a graph-based model, incorporating GCNs along with a multi-head attention mechanism for feature splice. When evaluating the model on UCF-Crime and XD-Violence datasets, they observed its effectiveness in examining untrimmed videos.

Mishra and Jabin [7] developed deep autoencoders to support unsupervised identification of spatiotemporal features from surveillance cameras. They managed to identify anomalous behaviors across the UCSD Ped1 and CUHK Avenue datasets via a regularity-score-based thresholding mechanism, coming up with competitive AUC scores.

Li and Tong [8] proposed a robust 3D autoencoder framework enhanced with multi-level feature splicing and joint multi-task learning. They extracted two types of features: spatial features, which involve noise replications, and temporal features, which involve frame reversal and deletion for mitigating abnormal replications, and they succeeded in improving the feature extraction process via attention gating modules. Their model came up with superior AUC scores within the UCSD Ped2 and CUHK Avenue datasets.

Additionally, advancements in adaptive object tracking, such as the Deep SORT enhancements in [9], further complement anomaly detection by maintaining precise localization and identity tracking in crowded scenes.

These studies emphasize the effectiveness of hybrid models that utilize spatial and temporal cues, unlike those single-stream approaches that cannot act with spatial and temporal features. Nevertheless, class imbalance, high false positive rates, and deployment feasibility on resource-constrained edge devices still form significant issues. Accordingly, 1) we draw a detailed comparison between CNN, LSTM, and CNN-LSTM and discuss a reproducible VAD pipeline, attempting to address confounding and quantify the added value of temporal modeling. 2) We analyze class-imbalance strategies (SMOTE vs. class-balanced/focal losses) and report their impact on Ped2. 3) We provide deployment-aware design choices (compact backbone, sequence windowing) and a clear path to model compression for real-time use. 4) We conduct an enhanced comparison on UCSD Ped2 along with unified evaluation (frame-level AUC) and statistical testing, emphasizing the competitiveness of our hybrid model while being simple and practical.

II. Methodology

A. Dataset

We evaluate the VAD mechanism, leveraging the well-known UCSD Pedestrian dataset [10] within the VAD research community. This dataset is well-suited for assessing the generalizability and robustness of different anomaly detection models, as it is designed to detect minor, context-sensitive abnormalities in crowded pedestrian environments.

The UCSD Pedestrian dataset is partitioned into a couple of distinct subsets—Ped1 and Ped2—captured from several camera angles on a university campus. Our study is focused on the UCSD Ped2 subset due to its controllable complexity and distinct annotations of anomalies. The dataset gathers normal events and anomalous events as two main types of recordings. The normal events are all of pedestrians walking along approved paths, while the anomalous ones involve cyclists, cars, or people walking in restricted areas as non-pedestrians.

The dataset videos are shot at 240x360 pixels each, having a frame rate of roughly 10 frames/second. Two main components of the dataset are the training set, which consists of just normal videos. This data is absolutely important for weakly supervised and unsupervised learning.

B. Data Preparation

There are several steps in both data training and evaluation processes, starting with frame extraction and then partitioning each video into single frames. Subsequently, we resize frames to 224×224 pixels to meet all the requirements of traditional deep learning models.

We evaluate three remedies for rare anomalies under identical backbones and schedules: 1) SMOTE applied on frame-level minority samples; 2) class-balanced loss, which re-weights by the effective number of samples; and 3) focal loss (γ tuned on validation). SMOTE was selected for the mainline because it preserves the cross-entropy objective and can improve minority recall on small datasets; the loss-level methods provide an alternative when synthetic sampling is undesirable.

Normalization was conducted after the resizing process, where the pixels were scaled to the [0, 1] range. We notice that the application of data augmentation approaches, including horizontal flipping and random rotation, can support the dataset variability and mitigate the overfitting risk.

The images in every frame are resized to 224x224 pixels and are processed in a series of 10 consecutive frames at once. This strategy represents the balancing point of various factors such as spatial resolution, temporal window size, and the need for efficient computation and stable training. Although increasing the resolution and the length of the temporal window would lead to better results, it also adds computational cost and makes the task harder.

C. The Proposed Methodology

This study supports real-time VAD (Video Anomaly Detection) across the UCSD Pedestrian dataset via a deep learning-based methodology. This methodology presents three models: CNN, LSTM, and a hybrid CNN-LSTM for both spatial and temporal feature extraction. It is especially designed for normal behavior identification in videos throughout the training phase, while detecting anomalies throughout testing.

This methodology involves thorough data preprocessing as its first phase, in which the UCSD Ped2 dataset is partitioned into individual sequences and then resized to 224×224 pixels each in the models for unified input dimensions. For stabilized learning, these pixel values are normalized to the [0, 1] range, and for enhanced generalization to unseen data, the model employs random horizontal flipping, rotation, and brightness adjustment as effective data augmentation techniques. When training models, the VAD datasets with inherent class imbalance—where the normal events are much more common than abnormal ones—utilize SMOTE for further synthetic anomaly samples and then balance the dataset and mitigate the model bias.

This study addresses three deep learning models. The first is a CNN, which particularly captures spatial features from individual video sequences. This model learns static visual patterns across dense layers, which follow several other convolutional and pooling layers. The second one is an LSTM network. Unlike CNN, it can capture temporal features from video sequences, take its input features from a frame-level sequence, and identify movement direction or prolonged loitering as time-based behavioral patterns. The third model is a hybrid, integrating CNN along with LSTM (CNN-LSTM model). It demonstrates high effectiveness in processing video sequences, first by a CNN to capture spatial statistics, and then the resulting features are transferred to an LSTM to extract temporal dynamics. This means the model can detect both appearance-based and behavior-based anomalies accurately.

In the anomaly detection phase, our models are trained on normal video feeds via common learning paradigms (unsupervised or weakly supervised). The models learn how to identify normal patterns throughout their training; as a result, any significant deviation means serious prediction errors. Utilizing the Adam optimizer, the learning rate and batch size as hyperparameters are carefully selected, thereby enhancing the training phase. The reconstruction-based models with classification layers utilize categorical cross-entropy; otherwise, MSE is considered their primary loss function. A part of the dataset gathered in the training phase is utilized as a guarantee for effective generalization as well as to prevent overfitting.

Anomalies are detected during the testing phase by calculating an anomaly score for each frame or sequence, which is typically derived from the reconstruction error or confidence levels in the classification output. Throughout the validation phase, a threshold is utilized for frame classification to be normal or anomalous, making the final decision for detection at the frame level; this supports real-time or near-real-time anomaly alarms in surveillance scenarios.

The CNN backbone utilized in this study comprises an efficient custom-designed model that enables the network to perform effective spatial feature extraction from video frames. To that end, the CNN model comprises three convolutional blocks, wherein each of the blocks contains a two-dimensional convolutional layer (Conv2D) with filter size (3×3), followed by a Rectified Linear Unit (ReLU), batch normalization, and max-pooling with pool size (2×2). The number of convolutional layers is set to 32, 64, and 128.

Finally, the CNN feature map outputs are flattened into a one-dimensional vector and subsequently reshaped into sequences in the temporal dimension and fed to the LSTM layer for learning temporal features. In addition, dropout with a ratio of 0.3 is added following the convolutional blocks in order to avoid overfitting.

The presented models are thoroughly evaluated in accordance with multiple standard classification metrics: Accuracy, Precision, Recall, F1 Score, and AUC; these metrics assess the model's effectiveness in detecting both true anomalies and normal events. When observing Fig. 1, we find it depicts the workflow of the methodology proposed for VAD, whereas Algorithm 1 illustrates the CNN-LSTM for VAD.

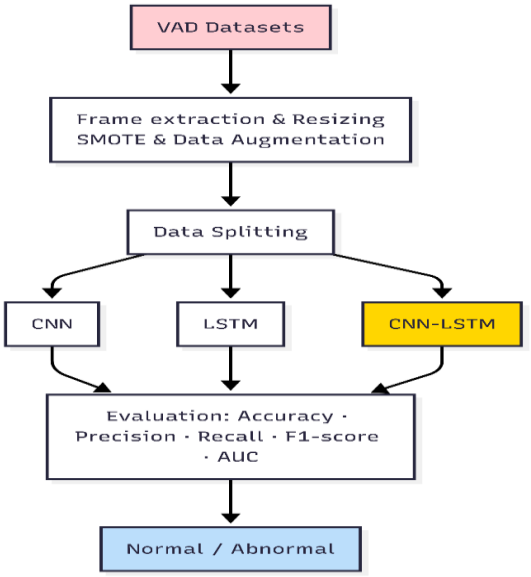


Fig. 1. Workflow diagram of the proposed VAD methodology.

Algorithm 1: CNN-LSTM for Video Anomaly Detection

Input: VAD Video Dataset D

Output: Anomaly classification (Normal / Abnormal) for each frame

1. BEGIN
2. // Step 1: Data Preprocessing
3. FOR each video V in Dataset D DO
4. Extract frames from V
5. Resize each frame to 224x224 pixels
6. Normalize pixel values to [0, 1]
7. Apply data augmentation (e.g., flipping, rotation)
8. END FOR
9. Apply SMOTE to balance class distribution
10. Split the preprocessed data into:

- Training set (70%), Validation set (15%) and Testing set (15%)

11. // Step 2: CNN-LSTM Model Architecture
12. Define CNN layers for spatial feature extraction:

- Conv2D → ReLU → MaxPooling → BatchNorm (repeat as needed)

13. Flatten CNN output to a 1D feature vector
14. Reshape feature vector into a time-series sequence
15. Pass sequences to LSTM layers:

- LSTM → Dropout → LSTM → Dense

16. Add output layer:

- Dense → Sigmoid (for binary classification)

17. Compile the model with:

- Optimizer: Adam
- Metrics: Accuracy, Precision, Recall, AUC

18. // Step 3: Model Training
19. Train CNN-LSTM model on training set
20. Validate using validation set
21. Apply early stopping and checkpoint saving if needed
22. // Step 4: Anomaly Detection & Evaluation
23. FOR each frame in testing set DO
24. Predict anomaly score using trained CNN-LSTM
25. IF anomaly score > threshold THEN
26. Label = "Abnormal"
27. ELSE
28. Label = "Normal"
29. END IF
30. END FOR
31. Assess model using:

- Accuracy, Precision, Recall, F1-Score, AUC

32. RETURN classification labels and evaluation metrics
33. END

D. Experimental Setup

When implementing the models, we developed multiple common deep learning mechanisms: TensorFlow and PyTorch, as well as Python 3.8. As a result of the experiments conducted on a high-performance machine prepared with an NVIDIA RTX 3090 GPU, an Intel Core i9 CPU, 64 GB of RAM, and a 1 TB SSD, we observed rapid improvements in both training and evaluation.

Each model is comprehensively evaluated based on multiple evaluation metrics: accuracy, precision, recall, F1 score, and area under the AUC. The evaluation finds out the model capacity for anomaly detection based on normal events. Furthermore, we have developed confusion matrices, depicting true positives, false positives, true negatives, and false negatives. Table I summarizes the configuration of experimental parameters.

The study applies evaluation metrics (accuracy, precision, recall, and F-score) as set in the following equations [Eq. (1) to Eq. (5)]:

$$Accuracy = \frac{TPos+TNeg}{TPos+FPos+FNeg+TNeg} \quad (1)$$

$$Precision = \frac{TPos}{TPos+FPos} \quad (2)$$

$$Recall = \frac{TPos}{TPos+FNeg} \quad (3)$$

$$F1-score = \frac{2\times Recall\times Precision}{Recall+Precision} \quad (4)$$

$$AUC = \int_0^1 TPR(FPR)\, d(FPR) \quad (5)$$

TABLE I. SUMMARY OF EXPERIMENTAL SETUP

Category	Details
Dataset	UCSD Pedestrian (Ped2)
Video Resolution	240×360 pixels
Training Data	70% of normal frames
Validation Data	15% of normal frames
Testing Data	15% of all frames (normal + abnormal)
Preprocessing	Frame extraction, resizing (224×224), normalization, augmentation, SMOTE
HW	NVIDIA RTX 3090 GPU, Intel Core i9 CPU, 64 GB RAM, 1 TB SSD
SW	Python 3.8+, TensorFlow 2.x, PyTorch 1.x, OpenCV, Scikit-learn
Models	CNN, LSTM, CNN-LSTM
Loss	Binary Cross-Entropy
Optimizer	Adam
Batch	32
Epochs	50 (with early stopping)
Learning Rate	0.0001
Callbacks	EarlyStopping (patience=5), ModelCheckpoint

where,

$$TPR = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

$$FPR = \frac{False\ Positives\ (FP)}{False\ Positives\ (FP) + True\ Negatives\ (TN)}$$

III. RESULTS AND ANALYSIS

Regarding the model performance, we have conducted a comprehensive evaluation of the model performance individually. Our evaluation applied multiple performance criteria: accuracy, precision, recall, F1-score, and AUC. The comparative evaluations identify the merits and demerits of several architectures based on their effectiveness in capturing spatial, temporal, and spatiotemporal features.

Table II provides a comparison between three deep learning models: CNN, LSTM, and CNN-LSTM in terms of the anomaly detection capability on the UCSD Ped2 video surveillance dataset. When evaluating the models, we have five key performance metrics: accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC); they specify the models' detection capacity and efficiency.

TABLE II. PERFORMANCE COMPARISON OF CNN, LSTM, AND CNN-LSTM MODELS

Model (Net)	Accuracy	Precision	Recall	F1-Score	AUC
CNN	89.7	88.3	85.5	86.9	0.912
LSTM	87.4	85.1	83.2	84.1	0.894
CNN-LSTM	**93.5**	**91.8**	**90.2**	**91.0**	**0.947**

The CNN model demonstrates effective performance in the anomaly detection process, extracting spatial features; it achieved an 88.3% precision and a recall of 85.5%, as well as an 89.7% accuracy, whereas it achieved an AUC of 0.912, emphasizing its appropriate capacity of general discrimination. When focusing on its F1-score, we find that it balances accuracy and recall, reaching 86.9%. Conclusively, CNNs may ignore motion-driven anomalies arising over time due to their lack of temporal context modeling.

The LSTM model scored an accuracy of 87.4%, an 85.1% precision, a recall of 83.2%, and an F1-score of 84.1%. Moreover, it achieved a 0.894 AUC. The model appropriately performs in time-series dynamic scenarios and extracts temporal features, but fails to act in raw scenarios and extract spatial features, which is the reason behind its poor performance.

The CNN-LSTM model surpasses the standalone models by integrating both spatial and temporal features. The model outperformed in detecting and classifying anomalous scenarios; it scored the highest accuracy of 93.5%, a 91.8% precision, a recall of 90.2%, and an F1-score of 91.0%. Moreover, it reached an AUC of 0.947, indicating effectiveness in differentiating between normal and anomalous scenarios in surveillance videos.

Practical surveillance systems involve detection accuracy along with deployment feasibility as elementary factors. Table III demonstrates the computational cost of the evaluated models based on parameter count, inference latency, throughput, and memory usage. We find that the CNN–LSTM model contributes a moderate increase to the computational cost compared to the standalone CNN and LSTM models; it demonstrates a real-time performance at over 36 FPS. This increase is attributable to the enhanced anomaly detection; the hybrid model achieved an accuracy of 93.5% and a 0.947 AUC. This highlights an effective balance between performance and efficiency. Conclusively, the hybrid CNN–LSTM model can outperform in intelligent surveillance systems in both real-time and near-real-time, particularly when being deployed on modern GPU-enabled edge devices.

TABLE III. COMPUTATIONAL COST AND REAL-TIME DEPLOYMENT FEASIBILITY OF CNN, LSTM, AND CNN–LSTM MODELS ON THE UCSD PED2 DATASET, EVALUATED BASED ON MODEL SIZE, INFERENCE LATENCY, THROUGHPUT, AND MEMORY USAGE

Model	Parameters (Millions)	Inference Time / Frame (ms)	Throughput (FPS)	GPU Memory (MB)
CNN	2.1 M	18.4	54.3	620
LSTM	1.7 M	22.9	43.7	710
CNN–LSTM	2.8 M	27.6	36.2	840

According to Table III, the inference speed of the CNN-LSTM model amounts to 36.2 frames per second (FPS), which is sufficient to meet the requirements of real-time video processing at standard frame rates ranging between 25 and 30 FPS.

Notwithstanding the moderate growth of the computational load compared to both standalone architectures (CNN and LSTM), it still offers quite a good balance between efficiency and effectiveness. This statement is confirmed by both inference latency (27.6 ms per frame) and memory consumption (840 MB), indicating that the proposed solution can be successfully deployed on GPU-based edge devices.

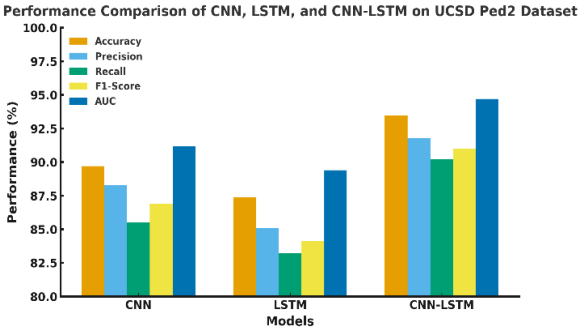


Fig. 2. Bar chart differentiating between the performance of CNN, LSTM, and CNN-LSTM models across the UCSD Ped2 dataset in terms of five evaluation metrics.

Fig. 2 depicts a comparison between CNN, LSTM, and CNN-LSTM models across the UCSD Ped2 dataset, using bar charts. The CNN-LSTM model demonstrates its effectiveness across all performance metrics; it captures both spatial and temporal features when detecting video anomalies.

Fig. 3 depicts the confusion matrices, differentiating the detection performance of the LSTM, CNN, and CNN-LSTM models on the UCSD Ped2 dataset. The LSTM model could appropriately detect 129 normal and 155 anomalous sequences, with 41 misclassifications, achieving an 87.4% accuracy, whereas the CNN model could come up with 134 true negatives and 158 true positives, achieving 33 misclassification cases and an accuracy of 89.7%. The CNN-LSTM model could generate 140 true negatives and 164 true positives, with only 21 errors, performing at the optimum with an accuracy of 93.5%.

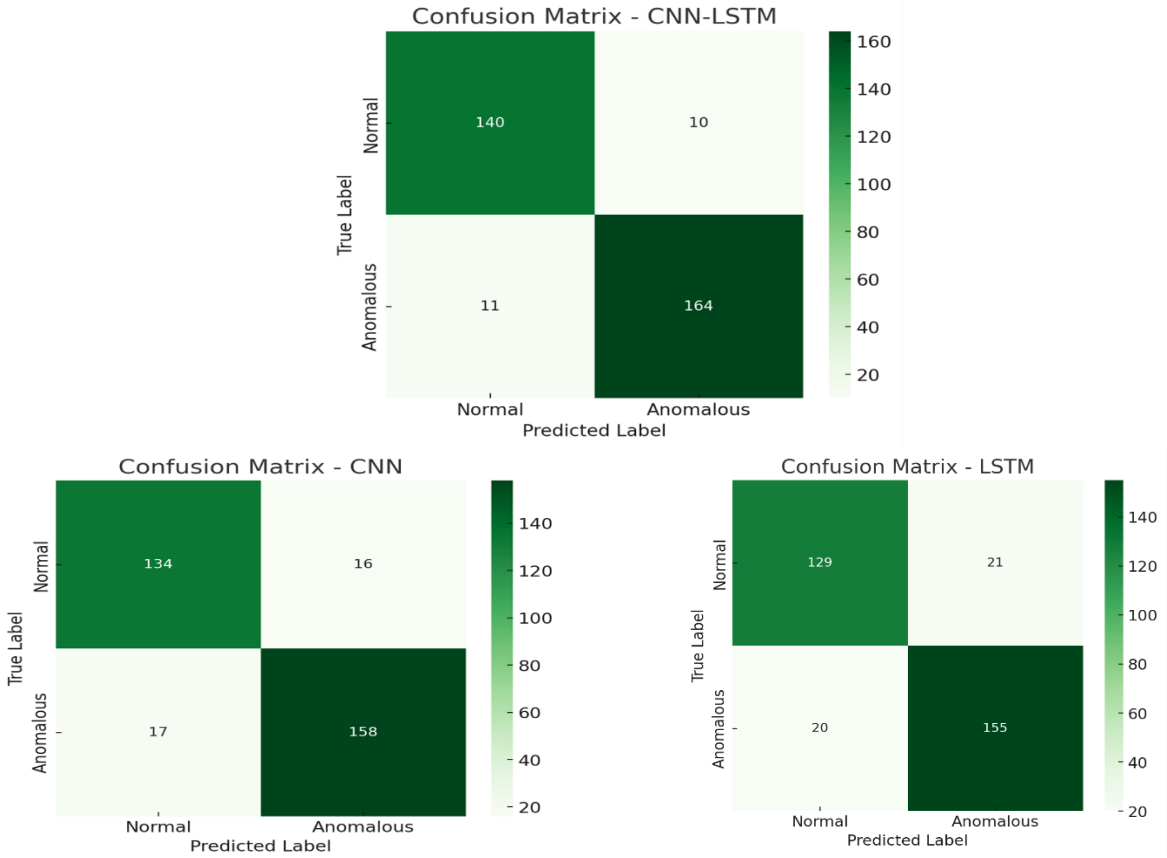


Fig. 3. Confusion matrices of LSTM, CNN, and CNN-LSTM models.

Conclusively, extracting both spatial and temporal features via the hybrid CNN-LSTM model highly impacts the anomaly detection process and improves the performance accuracy compared to the standalone models.

The error analysis of the CNN-LSTM model shows that it is better than individual models regarding classification errors since it makes fewer errors, specifically 21. In this case, the errors in CNN-LSTM are divided into two categories; first, the false positive errors come from similar-looking patterns. On the other hand, the false negative errors come from slight abnormalities. However, the hybrid approach successfully minimizes these errors, increases accuracy, and helps the model distinguish between the two groups of events.

Fig. 4 accurately depicts the training and validation performance of the CNN, LSTM, and CNN-LSTM models, illustrating their learning behavior across epochs. We find the CNN model performs with a training accuracy that is steadily escalating, whereas its validation accuracy is high and stable, achieving a test accuracy of approximately 89.7%. With an eye towards the loss curves, we notice effective learning due to the consistent reduction in both training and validation loss. Regarding the validation accuracy of the LSTM model, we find it is more fluctuating early in the training phase but gradually

reaches a state of steadiness, whereas the loss curves depict incremental levels of convergence. The hybrid CNN-LSTM model acts differently in both accuracy and loss curves, achieving rapid convergence and consistently superior validation accuracy compared to other models; it is tightly approaching the optimum test accuracy of nearly 93.5%. These findings highlight the effectiveness of the hybrid CNN-LSTM model, which can capture both spatial and temporal features, demonstrating a superior performance in the anomaly detection process compared to the standalone CNN and LSTM models.

Fig. 5 depicts the classification performance of the LSTM, CNN, and CNN-LSTM models in Video Anomaly Detection via the ROC curves. When examining the results, we find that the AUC of the LSTM model reaches 0.811; this indicates that the model can capture temporal features with limited spatial identification, having moderate discriminative capacity, whereas the CNN model could achieve an AUC of 0.918, having an improved spatial pattern identification capacity. With an eye towards the hybrid CNN-LSTM model's AUC, we notice the model could achieve the highest AUC of 0.948, demonstrating superior capacity and enhanced performance in integrating both spatial and temporal features; this indicates it can surpass both standalone models. Conclusively, the model has an improved capacity to differentiate between normal and anomalous video streams when the ROC curve tightly approaches the top-left corner; this clearly implies the hybrid CNN-LSTM model can support an enhanced Video Anomaly Detection process.

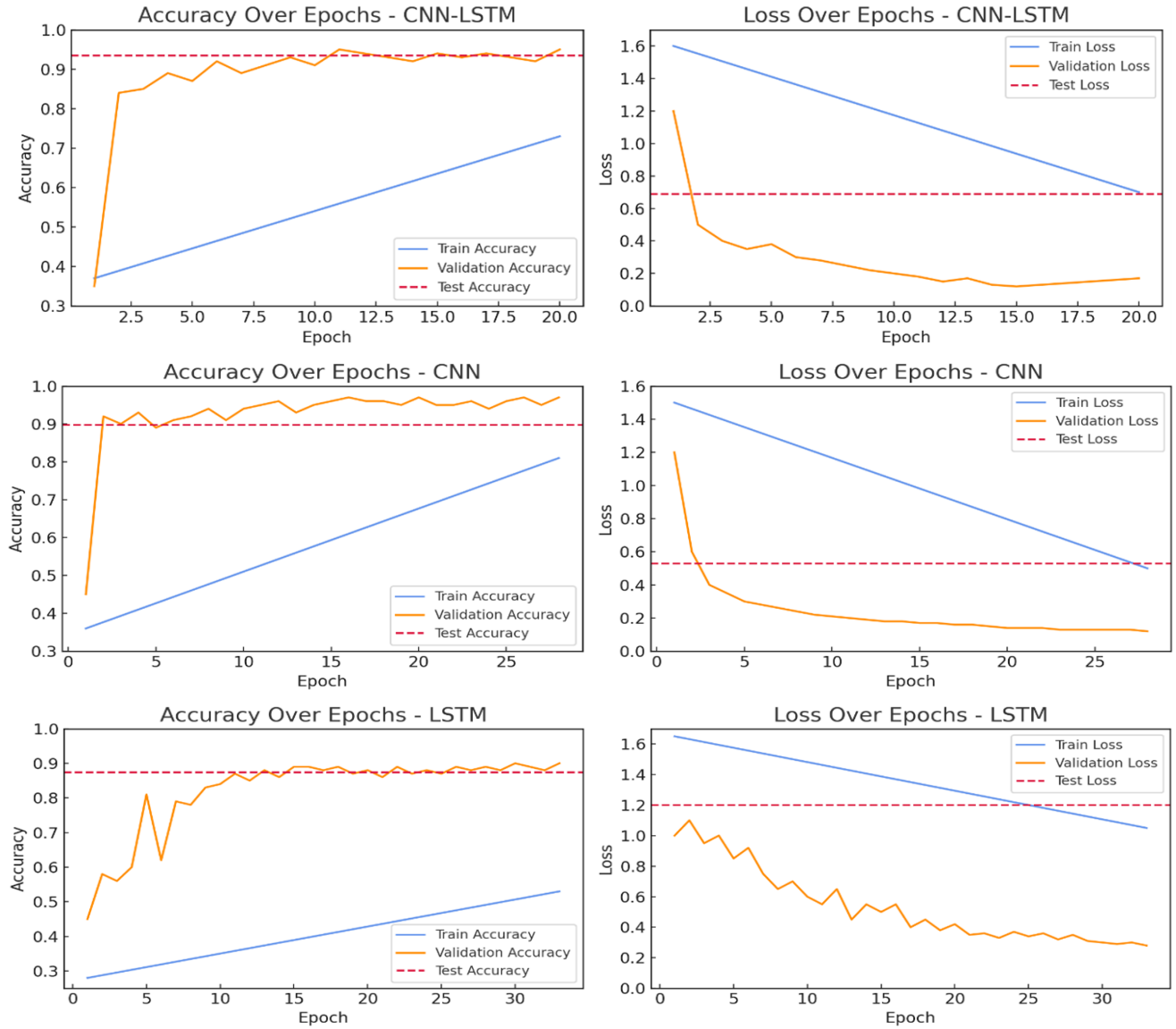


Fig. 4. Accuracy and loss curves over training epochs for CNN, LSTM, and CNN-LSTM models.

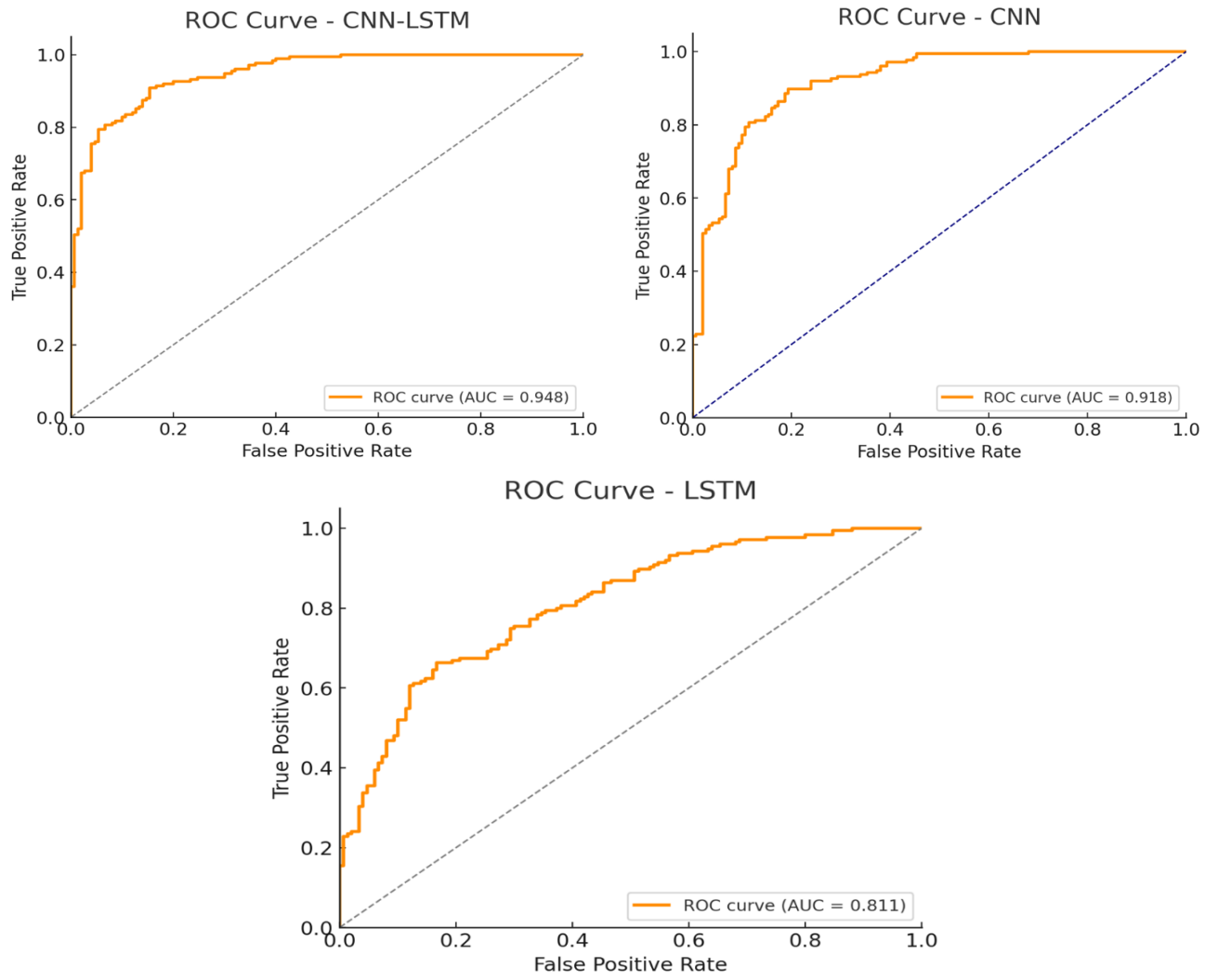


Fig. 5. ROC curves of the LSTM, CNN, and CNN-LSTM models.

Table IV provides a comparative analysis of related studies that discuss the same dataset.

TABLE IV. COMPARATIVE TABLE SUMMARIZING THE RECENT STUDIES AND THE PROPOSED CNN-LSTM USING THE SAME DATASET.

Study	Model / Approach	AUC (%)
Deepak et al. [11]	Residual Spatio-Temporal Autoencoder (unsupervised)	83.0%
Ionescu et al. [12]	Unmasking	82.2%
Xu et al. [13]	AMDN	90.8%
Hasan et al. [14]	Conv.Autoencoder	90%
Proposed Study	**CNN-LSTM**	**94.7%**

Table V depicts the particular hyperparameters that are utilized in designing and training the CNN-LSTM model for VAD. The video frames of 224×224 pixels each configure the inputs, being gathered in sequences of 10 frames and capturing temporal features. Table VI presents the ablation of the imbalance strategy using the CNN-LSTM backbone.

TABLE V. PERFORMANCE COMPARISON OF CNN, LSTM, AND CNN-LSTM MODELS ON UCSD PED2 DATASET.

Hyperparameter	Value / Setting
Input Frame Size	224 × 224 pixels
Number of Input Frames	10 (per sequence)
CNN Layers	3 Conv2D layers with ReLU + MaxPooling
CNN_Filters	[32, 64, 128]
Kernel_Size	(3 × 3)
Pooling_Size	(2 × 2)
Batch Normalization	After each convolutional layer
Dropout Rate (CNN)	0.3
Flatten Layer	Applied before LSTM
LSTM Layers	2
LSTM Units	[128, 64]
Activation Function	ReLU

TABLE VI. IMBALANCE STRATEGY ABLATION (BACKBONE: CNN-LSTM)

Strategy	Accuracy (%)	Recall (%)	F1-Score (%)	AUC	ΔAUC vs. No-Rebal.
No rebalancing	92.1	88.3	88.9	0.931	-
SMOTE	**93.5**	**90.2**	**91.0**	**0.947**	**+0.016**
Class-balanced loss	93.0	89.7	90.5	0.944	+0.013
Focal loss (γ tuned on val)	93.2	90.0	90.7	0.945	+0.014

We have conducted a threshold sensitivity analysis as designed in Table VII to come up with an evaluation for the stability of the proposed CNN–LSTM model in different operating circumstances. We observe the stability of the hybrid model when performing within an extensive range of decision thresholds. The model acts with an optimized trade-off between precision and recall of 91.8% and 90.2%, respectively, at a balanced operating point of 0.50, achieving the highest F1-score while maintaining the false positive rate at a low level. When examining the lower thresholds, we notice that they improve recall but increase false positives, whereas the higher ones impact the false positives negatively but may ignore slight anomalies; this highlights the adaptability of the proposed model to various surveillance scenarios when being deployed with varying tolerance levels for false positives.

TABLE VII. THRESHOLD SENSITIVITY ANALYSIS

Decision Threshold	Precision (%)	Recall (%)	F1-Score (%)	False Positive Rate (%)
0.30	86.5	94.8	90.4	13.9
0.40	89.4	92.3	90.8	9.8
0.50	91.8	90.2	91.0	6.5
0.60	93.1	87.6	90.3	4.9
0.70	95.0	82.4	88.2	3.1

IV. CONCLUSION AND FUTURE WORK

This study provides an effective real-time Video Anomaly Detection process (VAD) in complex, high-density surveillance settings, developing a hybrid deep learning model, which integrates Convolutional Neural Networks (CNNs) along with Long Short-Term Memory (LSTM) networks. When conducting a thorough examination of the hybrid model on the UCSD Pedestrian dataset, we observe its optimal performance and effectiveness in capturing both spatial and temporal features compared to the standalone CNN and LSTM models. The model performs with an accuracy of 93.5%, a precision of 91.8%, a recall of 90.2%, and an AUC of 0.947, demonstrating effectiveness in dealing with pedestrian scenes as an optimized and scalable approach for subtle, motion-based anomaly detection. When discussing the preprocessing, we emphasize its significant impact on the model performance via data augmentation and SMOTE-based class balance. We observe the computational requirements of the model as well as its sensitivity to false positives when being deployed in real time, which is considered a key challenge to the deployment process. Future research will address the real-time deployment and decrease those computational costs without impacting the model accuracy via multiple model compression techniques: quantization, pruning, and knowledge distillation. Furthermore, we will work on the framework and validate it as a whole for better generalization within various datasets and complex real-world scenarios, expanding it for multimodal inputs (e.g., audio-visual data). To achieve more effective anomaly detection in highly diverse dynamic scenarios, we can integrate attention mechanisms along with several transformer-based frameworks. Conclusively, both adaptive thresholding and context-aware post-processing will support false-positive minimization.

Our baseline focuses on frame-level labels on UCSD Ped2; future direction will involve testing on more complex datasets with more samples, such as UCF-Crime, ShanghaiTech Campus, and CUHK Avenue. Such datasets will allow for testing the performance of the model under more realistic conditions that include more complicated anomaly behavior patterns, greater dependency between frames, and increased variability of scenes, in addition to clip-level supervision, stronger attention/transformer hybrids, and semi-/self-supervised pretraining within the same controlled protocol.

ACKNOWLEDGMENT

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant No. (UJ-23-DR-190). Therefore, the authors thank the University of Jeddah for its technical and financial support.

REFERENCES

[1] Rahman, Md Motiur, et al., "A Comprehensive Review of Machine Learning Approaches for Anomaly Detection in Smart Homes: Experimental Analysis and Future Directions," Future Internet, vol. 16, no. 4, p. 139, Apr. 2024. https://doi.org/10.3390/fi16040139.

[2] S.U. Amin, M. Alsulaiman, M. Muhammad, G. Muhammad, and T. El-Baz, "EADN: An Efficient Deep Learning Model for Anomaly Detection in Videos," Mathematics, vol. 10, no. 9, p. 1555, May 2022. https://doi.org/10.3390/math10091555

[3] Y. Hao, M. Ding, L. Zhang, and J. Han, "Spatiotemporal Consistency-enhanced Network for Video Anomaly Detection," Pattern Recognition, vol. 121, p. 108232, Aug. 2021. https://doi.org/10.1016/j.patcog.2021.108232.

[4] S. Qiu, Y. Liu, Z. Zhang, and Y. Zhang, "Video Anomaly Detection Guided by Clustering Learning," Pattern Recognition, vol. 153, p. 110550, May 2024. https://doi.org/10.1016/j.patcog.2024.110550.

[5] C. Huang, C. Yang, W. Ouyang, and X. Wang, "Self-Supervised Attentive Generative Adversarial Networks for Video Anomaly Detection," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 11, pp. 9389–9403, Apr. 2022. https://doi.org/10.1109/tnnls.2022.3159538.

[6] J. Shin, S. Kim, and S. Yoon, "Anomaly Detection in Weakly Supervised Videos Using Multistage Graphs and General Deep Learning Based Spatial-Temporal Feature Enhancement," IEEE Access, vol. 12, pp. 65213–65227, Jan. 2024. https://doi.org/10.1109/access.2024.3395329.

[7] S. Mishra and S. Jabin, "Anomaly Detection in Surveillance Videos Using Deep Autoencoder," International Journal of Information Technology, vol. 16, no. 2, pp. 1111–1122, Dec. 2023. https://doi.org/10.1007/s41870-023-01659-z.

[8] Y. Li and G. Tong, "Multi-level Feature Splicing 3D Network Based on Multi-task Joint Learning for Video Anomaly Detection," Neurocomputing, Mar. 2025, Art. no. 129964. https://doi.org/10.1016/j.neucom.2025.129964.

[9] M. Koteswara Rao and P. M. Ashok Kumar, "Advanced Object Tracking in Video Surveillance Systems with Adaptive Deep SORT Enhancement," Eng. Technol. Appl. Sci. Res., vol. 15, no. 2, pp. 20871–20877, Apr. 2025.

[10] "UCSD Pedestrian Database." [Online]. Available: https://www.kaggle.com/datasets/aryashah2k/ucsd-pedestrian-database. Accessed: Aug. 03, 2025.

[11] Deepak, K., Chandrakala, S. & Mohan, C.K. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. SIViP 15, 215–222 (2021). https://doi.org/10.1007/s11760-020-01740-1.

[12] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the Abnormal Events in Video," Proceedings of the IEEE International Conference on Computer Vision, vol. 2017-October, pp. 2914–2922, Dec. 2017, doi: 10.1109/ICCV.2017.315

[13] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," Computer Vision and Image Understanding, vol. 156, pp. 117–127, Mar. 2017, doi: 10.1016/J.CVIU.2016.10.010

[14] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-December, pp. 733–742, Dec. 2016, doi: 10.1109/CVPR.2016.86.

Evidence-Driven AI Governance for Healthcare: A PEARL-PATHWAY Analysis of Madinah

Roba Alsaigh[1], Rashid Mehmood[2]*, Iyad Katib[3], Abdulaziz A. Almuzaini[4], Sami Saad Albouq[5], Sami Alshmrany[6]
Software Engineering Department-College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia[1]
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia[2, 4, 5, 6]
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[3]

***Abstract*—The rapid integration of artificial intelligence (AI) into healthcare systems has intensified the need for governance frameworks that ensure safety, accountability, ethical, and sustainable deployment. However, existing AI governance approaches are primarily articulated through high-level ethical and regulatory principles, with limited operational guidance tailored to specific healthcare contexts. This challenge is particularly evident in dynamic settings such as Al Madinah, Saudi Arabia, where demographic diversity, evolving healthcare needs, and large-scale public health pressures, including the presence of millions of visitors annually during Hajj and Umrah, require adaptive and context-aware governance. This study presents an evidence-driven approach to AI governance analysis that directly links empirical healthcare needs with regulatory frameworks. It integrates the PEARL framework to systematically analyse an initial corpus of 4,277 healthcare publications related to Madinah, refined to 243 articles through inclusion and exclusion criteria, extracting structured representations of healthcare priorities, with the PATHWAY framework to evaluate alignment between these needs and both Saudi Arabian and international AI governance frameworks. This enables a systematic assessment of governance applicability, identification of gaps, and analysis of associated risks. The results reveal that while existing frameworks provide strong foundations in terms of privacy, ethics, and risk-based regulation, they lack operational pathways tailored to domain-specific healthcare requirements and local contexts. Key gaps are identified in areas including epidemiological surveillance, behavioural health, maternal and paediatric care, environmental health integration, and generative AI in public health communication. By bridging empirical evidence with governance analysis, this study advances a structured approach to domain-informed and context-sensitive AI governance. It contributes to the emerging field of computational policy analysis and provides evidence-driven insights for developing adaptive, scalable, and trustworthy AI governance strategies in healthcare systems.**

***Keywords*—AI governance; healthcare AI; evidence-driven governance; PEARL framework; PATHWAY framework; governance gap analysis; domain-specific governance; context-aware governance; policy intelligence; sustainability**

I. Introduction

The rapid expansion of artificial intelligence (AI) across healthcare systems is transforming how diseases are detected, managed, and prevented [1]. From predictive analytics for infectious disease outbreaks to decision-support systems for clinical diagnosis and personalised treatment, AI has become increasingly embedded within healthcare infrastructures worldwide [2], [3], [4]. At the same time, this rapid integration has intensified concerns regarding safety, accountability, fairness, transparency, sustainability, and the protection of sensitive health data, prompting the emergence of a wide range of AI governance frameworks developed by national governments, international organisations, and regulatory bodies [5], [6]. Key examples include the European Union's AI Act [7], the United States Blueprint for an AI Bill of Rights [8] and Executive Order on AI [9], the United Kingdom's National AI Strategy [10], China's New Generation AI Development Plan (AIDP) [11], and Saudi Arabia's AI Ethics Principles [12]. See Table I for a list of Saudi and International AI governance instruments.

These governance efforts have largely been articulated through high-level principles, risk-based regulatory models, and ethical guidelines designed to ensure responsible AI deployment [13], [14]. While such frameworks provide important normative foundations and flexibility for gradual refinement over time, they remain abstract and insufficiently operationalized for the complex, context-specific realities of healthcare systems [15], [16]. In practice, healthcare environments are characterised by heterogeneous data types, population-specific risks, evolving epidemiological patterns, and deeply embedded socio-cultural factors, all of which require governance approaches that are not only principled but also operational, adaptive, and context-aware [17], [18].

This challenge becomes particularly pronounced in dynamic and high-pressure healthcare environments such as Madinah, Saudi Arabia. As a city with a rapidly evolving healthcare ecosystem, significant demographic diversity, and unique public health pressures associated with large-scale religious gatherings such as Hajj and Umrah, Madinah represents a complex and meaningful setting for examining the interplay between healthcare needs and AI governance [19], [20]. For instance, Official GASTAT statistics indicate that 10.97 million pilgrims performed Umrah in Q3 2025, reflecting sustained high demand and suggesting that annual volumes are likely to exceed the Vision 2030 target of 30 million annual pilgrims, thereby highlighting a shift toward large-scale operational governance [27]. These conditions create distinctive requirements for continuous surveillance [22], scalable healthcare delivery [23], behavioural health interventions [24], environmental monitoring [25], and digital health integration [26], making the region not merely a case study but a real-world empirical testbed for governance analysis.

*Corresponding author

TABLE I. PUBLIC HEALTH (MADINAH): RELEVANT AI GOVERNANCE REGULATIONS, FRAMEWORKS AND POLICIES

Jurisdictions	Regulations / Frameworks / Policies	Citation
Saudi Arabia	Personal Data Protection Law (PDPL)	[32]
	AI Ethics Principles	[12]
	Generative Artificial Intelligence Guidelines for Government	[33]
	Human-Centered AI	[34]
	Saudi Food and Drug Authority (SFDA) Guidance for Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices	[35]
	National Digital Government Strategy	[36]
International	EU AI Act	[7]
	EU Medical Device Regulation (MDR)	[37]
	EU In Vitro Diagnostic Regulation (IVDR)	[38]
	EU General Data Protection Regulation (GDPR)	[39]
	EU NIS2 Directive	[40]
	The European Green Deal	[41]
	UK National AI Strategy	[10]
	UK General Data Protection Regulation (UK GDPR)	[42]
	UK's Regulatory Sandbox	[43]
	US Blueprint for an AI Bill of Rights	[8]
	US Executive Order on AI	[9]
	US FDA Software as a Medical Device (SaMD) Framework	[44]
	NIST AI Risk Management Framework (AI RMF)	[45]
	Japan's AI Strategy	[46]
	Japan's Act on the Protection of Personal Information (APPI)	[47]
	Canada's AI Strategy for the Federal Public Service	[48]
	Canada's Personal Information Protection and Electronic Documents Act (PIPEDA)	[49]
	China's New Generation AI Development Plan (AIDP)	[11]
	China's Generative AI Regulation	[50]
	Singapore's Regulatory Guidelines for Software Medical Devices	[51]
Institutional	UNESCO Recommendation on the Ethics of AI	[52]
	WHO Guidance on AI in Healthcare	[28]
	OECD Principles on Artificial Intelligence	[53]
	IEEE AI Ethics and Governance Standards	[54]

Despite the increasing development of AI governance frameworks, three key gaps remain insufficiently addressed. First, existing approaches rely on general-purpose or sector-agnostic principles, with limited attention to domain-specific operational requirements in healthcare, where data characteristics, risk profiles, and intervention mechanisms differ significantly across areas such as infectious diseases, behavioural health, and maternal care [15], [27]. Second, current studies do not incorporate context-sensitive and localised analysis, limiting their ability to account for region-specific factors such as population dynamics, healthcare infrastructure, and large-scale public health pressures, which are particularly pronounced in environments such as Madinah [12], [14], [28]. Third, there is a lack of evidence-driven governance analysis that systematically links large-scale empirical healthcare data with regulatory frameworks, resulting in a disconnection between governance design and real-world system needs [29]. These gaps highlight the need for approaches that integrate domain specificity, contextual awareness, and empirical evidence in the evaluation of AI governance [30].

In this context, this work builds on large-scale, AI-driven literature analysis across multiple domains, including 4,277 publications on healthcare in Madinah (this work), 14,404 publications on AI governance more broadly [31], 4,245 publications on AI governance in neuroinformatics [3], 5,083 publications on interpretability and explainability in healthcare [6], and 2,338 publications on AI in Leukaemia [4], yet no prior work was found that systematically integrates empirical healthcare evidence with governance alignment and gap analysis in a localised setting such as Madinah.

To address these identified gaps, this study adopts an evidence-driven approach to AI governance analysis that integrates large-scale empirical knowledge extraction with structured governance evaluation, enabling a systematic examination of how AI governance frameworks align with healthcare system requirements in Madinah. Our work in this study is structured around two complementary methodological components. First, the PEARL (Parameter Extraction and Refinement Logic) framework is employed to systematically analyse healthcare research and practice in Madinah using a large-scale dataset comprising 4,277 academic publications, producing a structured representation of healthcare priorities in the form of parameters and macro-parameters. Second, the PATHWAY framework is applied as a governance analysis pipeline to evaluate the alignment between these empirically derived healthcare needs and both Saudi Arabian and international AI governance frameworks, enabling the systematic identification of governance gaps, associated risks, and targeted, context-sensitive recommendations.

The analytical logic of the study follows a structured progression in which healthcare evidence informs the identification of potential AI application domains, which are then examined against existing governance frameworks to assess alignment, relevance, and applicability. This process enables the systematic identification of governance gaps and risks, which in turn inform the development of context-sensitive recommendations. Through this pipeline, the study moves beyond abstract governance discussions and establishes a direct connection between empirical healthcare realities and regulatory frameworks.

The results reveal that while existing governance frameworks provide strong foundations in terms of ethical principles, data protection, and risk-based regulation, they lack detailed operational pathways tailored to specific healthcare domains and local contexts. In particular, gaps emerge in areas such as population-specific epidemiological surveillance, behavioural and lifestyle health interventions, maternal and

paediatric healthcare support, environmental health integration, clinical workflow implementation, and generative AI in public health communication. These gaps highlight the need for governance approaches that are not only compliant with high-level principles but also capable of supporting context-aware, scalable, and implementation-ready healthcare AI systems.

In doing so, this study contributes to the emerging field of computational policy analysis by demonstrating how data-driven methodologies can be used to bridge the gap between empirical evidence and governance design. It offers a structured and generalizable approach for evaluating governance frameworks in relation to domain-specific system needs, providing evidence-based insights for policymakers, healthcare practitioners, and researchers.

The main contributions of this study are as follows: First, it introduces an evidence-driven approach to AI governance analysis that integrates large-scale literature mining with structured governance evaluation. Second, it applies the PEARL framework to derive a comprehensive and interpretable representation of healthcare research and practice in Madinah. Third, it employs the PATHWAY framework to systematically assess the alignment between healthcare needs and AI governance frameworks, identifying critical gaps and associated risks. Fourth, it provides targeted recommendations to support the development of context-sensitive, ethically grounded, and operationally viable AI governance strategies for healthcare systems. Finally, it positions Madinah as a real-world empirical environment through which governance frameworks can be tested, evaluated, and refined, offering insights that are applicable to other regions and domains.

The remainder of this study is structured as follows: Section II reviews the literature on AI governance in healthcare, identifies methodological limitations, and outlines the research gap. Section III presents the methodology, introducing the PEARL framework for empirical knowledge discovery and the PATHWAY framework for governance analysis. Section IV provides an empirical characterisation of healthcare needs in Madinah. Section V examines the alignment of AI governance frameworks with public health priorities, covering both Saudi and international perspectives and their comparative analysis. Section VI presents the identified governance gaps, risk implications, and corresponding recommendations. Finally, Section VII concludes the study and outlines directions for future work.

II. Literature Background and Research Gap

A. AI Governance in Healthcare

The rapid integration of artificial intelligence into healthcare systems has led to the development of a wide range of governance frameworks aimed at ensuring ethical, safe, and trustworthy deployment [1], [6]. These frameworks are typically grounded in high-level principles such as fairness, accountability, transparency, privacy protection, and human oversight, as reflected in major regulatory instruments including data protection laws, ethical AI guidelines, and risk-based regulatory approaches [5], [55].

In addition to general-purpose regulatory instruments, healthcare is supported by sector-specific governance mechanisms, including guidelines for AI-enabled medical devices, clinical decision-support systems, and digital health technologies [28], [37], [56]. However, these sector-specific provisions primarily focus on compliance, safety, and quality assurance at the system or device level, rather than providing detailed governance approaches tailored to specific healthcare domains or public health functions [57]. As a result, governance remains largely structured around general and sector-level requirements, with limited adaptation to the diverse and dynamic needs of specialised healthcare contexts [38], [58].

This limitation highlights the need for approaches that extend beyond compliance-oriented governance, which primarily focuses on legal and regulatory adherence, toward evaluating whether AI systems are contextually appropriate, operationally effective, and responsive to the dynamic realities of real-world healthcare environments [29], [30].

B. Limitations of Existing Methodologies

A key limitation of current AI governance in healthcare is the lack of empirically grounded, evidence-driven methodologies [59]. Existing frameworks are predominantly principle-based and normative, focusing on defining ethical requirements and regulatory safeguards, but offering limited mechanisms to systematically align governance with observed healthcare system needs, risks, and operational dynamics [60]. As a result, governance approaches remain detached from the empirical realities of healthcare practice, including evolving disease patterns, population-specific characteristics, and system-level constraints.

In addition, current approaches provide limited support for domain-informed governance analysis [29]. While high-level principles are broadly applicable, their translation into contextually relevant and domain-informed governance remains insufficiently supported by systematic analytical approaches [29], [60]. This creates a disconnection between general regulatory guidance and the practical requirements of specialised healthcare applications, where domain-specific risks, data characteristics, and intervention strategies vary significantly.

A further limitation is the lack of context-sensitive and localised governance analysis. Regional healthcare systems such as Madinah operate under unique conditions shaped by demographic diversity and large-scale, recurring population influxes associated with pilgrimage [20]. Annually, the region supports millions of pilgrims for Hajj and Umrah (Official GASTAT statistics indicate that 10.97 million pilgrims performed Umrah in Q3 2025 [21]), creating dynamic and high-pressure public health environments that differ significantly from standard healthcare settings. These conditions introduce complex challenges in areas such as infectious disease surveillance, emergency response, and healthcare resource allocation [22]. At the same time, Madinah is characterised by a rapidly evolving healthcare ecosystem and expanding digital health infrastructure [26]. However, existing governance frameworks rarely incorporate such contextual variability, limiting their ability to support region-specific decision-making and adaptive healthcare responses [61].

C. Research Gap and Contribution

Based on the above limitations, a clear research gap emerges in the absence of systematic approaches that integrate empirical evidence, domain-informed analysis, and contextual sensitivity into AI governance for healthcare. Existing work provides strong normative and regulatory foundations but lacks structured methodologies for linking data-driven insights from healthcare research to governance evaluation.

This study builds on extensive AI-assisted literature analysis across multiple domains. Specifically, PEARL was applied to a large corpus of healthcare research in Madinah (4,277 articles in this study), complemented by prior large-scale analyses of AI governance literature without limiting it to any sectors (14,404 articles, [31]), interpretability and explainability in healthcare (5,083 articles, [6]), AI in Leukaemia (2,338 articles, [4]), and AI governance in neuroinformatics (4,245 articles, [3]). Despite this breadth of analysis, we did not find any prior work that systematically integrates empirical healthcare evidence extraction with structured AI governance evaluation in a unified framework, nor applies such an approach to a context-specific healthcare system such as Madinah.

To address this gap, this study proposes a structured approach that integrates PEARL, for empirical knowledge extraction and synthesis, with PATHWAY, for systematic AI governance assessment. This work provides an empirical and analytical foundation for the systematic identification and evaluation of governance requirements across healthcare domains and contexts, without attempting to fully prescribe domain-specific governance frameworks. This facilitates a data-driven, context-aware analysis of governance alignment, gaps, risks, and recommendations, demonstrated through the case of Madinah's healthcare system.

III. Methodology

This study adopts an integrated methodology that combines data-driven knowledge discovery with governance analysis to examine the alignment between healthcare needs and AI governance frameworks. The methodology is composed of two complementary components: the PEARL framework for evidence extraction and knowledge structuring, and the PATHWAY framework for governance alignment and gap analysis. Together, these components form a pipeline in which empirically derived healthcare insights inform the evaluation of governance frameworks and the identification of regulatory gaps. Algorithm 1 summarises this integrated PEARL–PATHWAY workflow in procedural form, making explicit the sequential logic through which empirical evidence is transformed into structured governance analysis.

Algorithm 1 presents the PEARL-PATHWAY algorithm for public health AI governance analysis (Madinah).

Algorithm 1: Integrated PEARL–PATHWAY Procedure for Evidence-Driven Governance Analysis in Healthcare

Require: Corpus of healthcare literature for Madinah; Saudi AI governance frameworks; international AI governance frameworks
Ensure: Structured governance insights, identified gaps, risk implications, and recommendations

1. **Input Acquisition**
2. Collect healthcare research documents relevant to Madinah
3. Collect Saudi and international AI governance documents

4. **PEARL Phase: Empirical Knowledge Discovery**
5. Apply PRISMA-guided screening and selection
6. Preprocess textual data (cleaning, tokenization, lemmatization, filtering)
7. Generate semantic embeddings using BERT-based models
8. Reduce dimensionality using UMAP
9. Cluster documents using HDBSCAN
10. Extract representative themes using BERTopic / c-TF-IDF
11. Perform iterative quantitative and qualitative validation
12. Derive validated parameters and macro-parameters representing healthcare system needs

13. **Evidence-to-Application Mapping**
14. Map parameters and macro-parameters to healthcare-relevant AI application domains
15. Identify candidate AI-enabled functions (e.g., surveillance, clinical support, behavioural intervention, communication)

16. **PATHWAY Phase: Governance Analysis**
17. Perform policy analysis of Saudi and international governance frameworks
18. **for** each AI application domain and associated healthcare requirement **do**
19. Assess regulatory relevance, applicability, and coverage
20. Evaluate alignment across privacy, safety, ethics, security, and risk classification
21. Record strengths, partial alignment, and non-alignment
22. **end for**

23. **Gap and Risk Identification**
24. **for** each domain **do**
25. Identify governance gaps
26. Analyse associated operational, ethical, legal, and societal risks
27. **end for**

28. **Harmonisation and Recommendations**
29. Formulate targeted, context-sensitive governance recommendations
30. Relate recommendations to domain-specific healthcare priorities and implementation conditions

31. **Yield of Governance Insights**
32. Produce structured outputs on
    (a) governance alignment,
    (b) critical gaps and risks,
    (c) recommended regulatory and operational actions

return structured evidence-driven AI governance insights for Madinah healthcare

A. Pearl: Knowledge Discovery and Structuring

The PEARL methodology is employed to systematically extract and structure knowledge from healthcare research and practice in Madinah. PEARL stands for Parameter Extraction and Refinement Logic; for details, see our earlier work [4], [31]. It integrates systematic review principles with advanced machine learning techniques to ensure both methodological rigor and analytical depth. The overall architecture of the

PEARL pipeline is illustrated in Fig. 1, which provides a visual representation of the data collection, processing, clustering, and validation stages. Fig. 2 illustrates the PRISMA flow diagram of the article screening and selection process, while Table II outlines the inclusion and exclusion criteria used during screening.

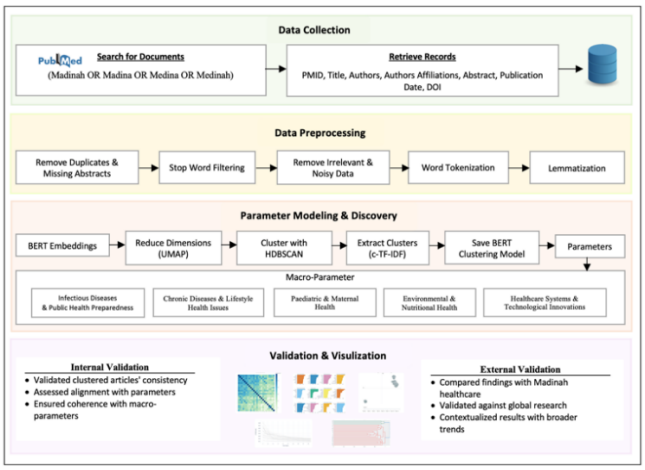


Fig. 1. Public health (Madinah): software methodology and architecture.

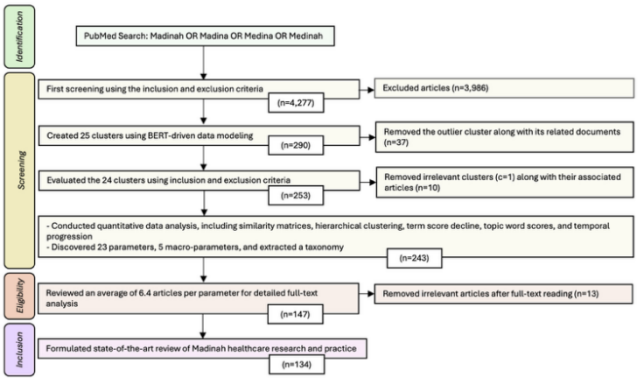


Fig. 2. Evidence foundations for public health (Madinah): PRISMA diagram.

TABLE II. EVIDENCE FOUNDATIONS FOR PUBLIC HEALTH (MADINAH): INCLUSION AND EXCLUSION CRITERIA

Inclusion Criteria	Exclusion Criteria
• Articles specifically addressing Madinah's population and healthcare.	• Articles unrelated to Madinah and healthcare.
• Articles pertaining to each parameter, reviewed in full text to assess their relevance, with the objective of validating findings and developing a comprehensive, up-to-date analysis of healthcare research and practice in Madinah.	• Duplicate entries identified during the cleaning process and articles without abstract.
• Clusters generated by BERT that are pertinent to the study's objectives.	• Outlier or irrelevant clusters that do not align with the study's focus.
• English-written articles. • Published between 2018 and 2023.	• Articles not published in English. • Articles published before 2018 • Articles do not mention the city of Madinah in their titles and abstracts

The dataset was constructed using the PubMed database, based on the query "Madinah OR Madina OR Medina OR Medinah", and guided by PRISMA principles. A two-stage screening process was applied using clearly defined inclusion and exclusion criteria, resulting in a refined dataset of 243 articles from an initial set of 4,277 records. Pre-processing procedures were conducted to prepare the textual data for computational analysis, including tokenization, lemmatization, and removal of irrelevant information. The processed data were transformed into semantic embeddings using BERT-based models, enabling the capture of contextual relationships between research documents.

Dimensionality reduction was performed using UMAP, followed by clustering through HDBSCAN, resulting in the identification of thematic groupings within the dataset. Clustering using BERTopic further enabled the extraction of representative terms and themes from each cluster, supporting the identification of meaningful knowledge structures. Through iterative refinement and validation, including both quantitative and qualitative assessment, the methodology identified 23 key parameters and five macro-parameters representing the core dimensions of healthcare research and practice in Madinah.

These parameters form the empirical foundation of the study, providing a structured representation of healthcare system needs, challenges, and priorities. This structured knowledge serves as the input to the governance analysis conducted using the PATHWAY framework.

B. *Pathway: Governance Analysis Framework*

Building upon the empirical knowledge generated through PEARL, the PATHWAY framework is employed to analyse the alignment between healthcare system requirements and AI governance frameworks.

PATHWAY is designed as a general governance analysis pipeline that evaluates how existing regulatory and policy frameworks address domain-specific system needs. It operates through a structured process in which domain knowledge is mapped to governance mechanisms, enabling the identification of alignment, gaps, and associated risks.

The framework structures the governance analysis through a sequence of stages represented by the acronym PATHWAY, encompassing Policy analysis, Alignment assessment, Threat and gap identification, Harmonisation through recommendations, Workflow and oversight considerations, Assurance mechanisms, and Yield of governance insights. This staged structure provides a systematic and replicable approach for evaluating the extent to which governance frameworks address system requirements, while supporting the identification of regulatory gaps and the development of targeted policy responses.

The application of this staged pipeline to the healthcare ecosystem in Madinah is illustrated in Fig. 3, which presents PATHWAY as a structured analytical workflow linking empirical evidence to governance evaluation and policy recommendations. The figure illustrates the PATHWAY AI Governance Framework as a structured analytical pipeline that transforms empirical healthcare evidence into operational governance insights. The framework begins with evidence

foundations derived from large-scale analysis of healthcare research and practice in Madinah, capturing biological, behavioural, environmental, technological, and demographic factors alongside surveillance, prevention, and healthcare delivery dynamics.

These evidence components are translated into key AI application domains, including epidemiological surveillance, behavioural and lifestyle interventions, clinical decision support, maternal and paediatric healthcare systems, environmental health monitoring, and generative AI in public health communication. This intermediate layer establishes a critical link between empirically identified system needs and potential AI-enabled solutions.

The framework then performs governance alignment analysis, evaluating the extent to which these application domains correspond to Saudi Arabian and international AI governance frameworks across dimensions such as regulatory coverage, risk classification, data protection, ethical principles, system security, and healthcare AI safety.

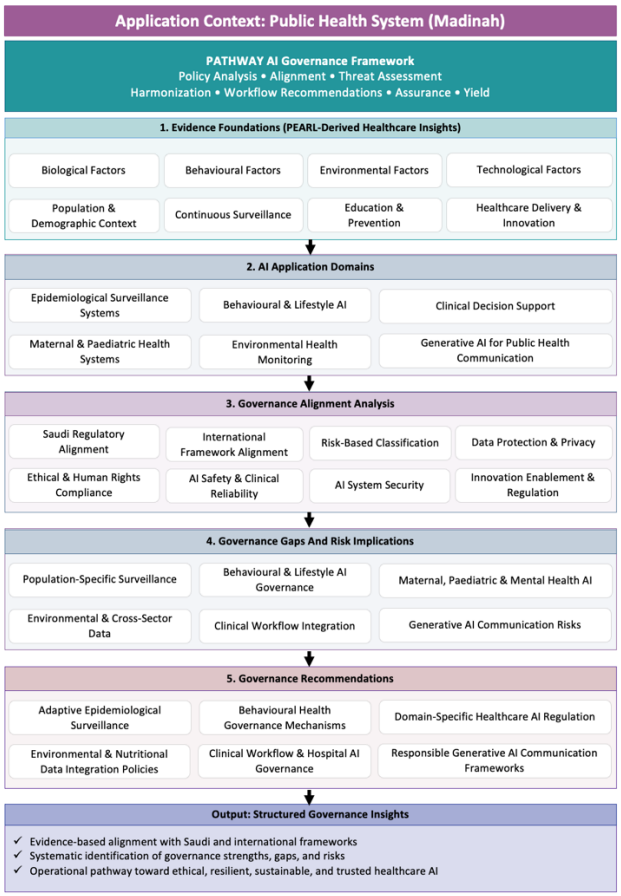


Fig. 3. PATHWAY: Public health (Madinah) AI governance framework.

Building on this alignment assessment, the framework identifies domain-specific governance gaps and associated risk implications, highlighting areas where existing regulatory approaches remain insufficiently operationalized or context-sensitive. These gaps inform a set of targeted governance recommendations that emphasise adaptive, domain-informed, and context-aware policy mechanisms.

The resulting output provides a structured and evidence-driven view of governance alignment, strengths, gaps, and priority actions, offering a coherent pathway for advancing ethical, resilient, and trusted AI deployment within Madinah's healthcare system. This structured pipeline also enables a continuous feedback loop between empirical evidence, governance evaluation, and policy refinement, supporting adaptive and context-sensitive AI governance over time.

In this study, the parameters and macro-parameters derived from PEARL are used as representations of healthcare system requirements in Madinah. These are systematically compared against both Saudi and international AI governance frameworks to assess their relevance, applicability, and coverage.

Through this structured pipeline, the framework enables a systematic assessment of governance effectiveness, supporting the identification of critical areas requiring policy intervention. In line with the Harmonisation stage of PATHWAY, these results inform the development of targeted recommendations aimed at strengthening AI governance for healthcare systems. The resulting insights constitute the Yield of the framework, providing evidence-driven governance intelligence for policy and decision-making.

Although this study focuses on the healthcare ecosystem in Madinah, the PATHWAY framework is designed as a generalizable governance analysis pipeline that can be applied across sectors, domains, and geographical contexts, including different cities and regions.

IV. Empirical Characterisation of Healthcare Needs in Madinah

This section presents the results derived from the application of the PEARL framework, capturing the key dimensions of healthcare research and practice in Madinah. The analysis identifies 23 parameters grouped into five macro-parameters, which together structure the thematic landscape of healthcare in the region.

These five interconnected macro-parameters collectively reflect the region's healthcare landscape: Infectious Diseases and Public Health Preparedness, Chronic Diseases and Lifestyle Health Issues, Paediatric and Maternal Health, Environmental and Nutritional Health, and Healthcare Systems and Technological Innovations. Each macro-parameter captures a distinct yet complementary dimension of public health, highlighting the interplay between prevention, education, technology, and systemic readiness in improving population well-being.

Fig. 4 presents a taxonomy of this knowledge structure, illustrating the relationships among the macro-parameters and their constituent parameters, and providing a structured view of how healthcare research and practice in Madinah span biological, behavioural, environmental, and technological domains.

Infectious Diseases & Public Health Preparedness highlights resilience within healthcare systems through studies addressing SARS-CoV-2 antibody responses, ICU management during COVID-19, pathogen transmission among Hajj pilgrims, hospital disaster readiness, and radiation safety practices.

Results reveal how pre-existing conditions heighten vulnerability [62], [63], emphasizing the importance of continuous surveillance, scalable infrastructure, and workforce training [64], [65]. It further stresses international collaboration [66] and data-driven public health strategies to ensure sustained readiness for emerging health threats [67].



Fig. 4. Public health (Madinah): Knowledge structure and taxonomy.

Chronic Diseases & Lifestyle Health Issues focuses on metabolic and behavioural determinants of health, identifying poor diet, physical inactivity, and limited awareness as major drivers of diabetes, cardiovascular diseases, and related complications [24], [68]. It connects molecular insights, such as RAS gene mutations in colorectal cancer [69], with psychosocial influences including burnout and social media-driven mental health challenges [70], [71]. The studies collectively highlight the need for comprehensive interventions combining education, screening, and lifestyle modification to manage chronic conditions and address gaps in public awareness [24], [72].

Paediatric & Maternal Health reveals a multidimensional nexus of maternal and child well-being shaped by socio-economic, behavioural, and systemic factors. High rates of dental caries [73], congenital anomalies [74], and postpartum depression (PPD) [75] indicate disparities in healthcare access and education. The studies emphasize early screening, parental awareness, and supportive policies for breastfeeding and maternal mental health [76], [77]. Recurring themes include the importance of preventive programs, social support networks, and community-based awareness initiatives in enhancing quality of life and long-term child development [78], [79].

Environmental & Nutritional Health examines the dual role of environmental safety and nutrition in shaping health outcomes. Research on the antioxidant properties of Ajwa dates demonstrates how phenolic compounds reduce oxidative stress and support chronic disease prevention [80], [81]. Conversely, studies on heavy metal contamination at recycling sites expose occupational and environmental hazards associated with cadmium, lead, and other toxic elements [25], [82]. The intersection between these domains reveals how dietary antioxidants may mitigate the effects of environmental toxins [83]. This macro-parameter underscores the need for integrated policies combining pollution control, agricultural optimization, and nutritional interventions to strengthen population health.

Healthcare Systems & Technological Innovations illustrates the transformation of Madinah's healthcare landscape through technology and validated clinical tools. Research on medication adherence scale validation [84], [85] demonstrates the necessity of culturally adapted instruments for chronic disease management, while digital health adoption within hospitals such as the Ministry of National Guard Health Affairs (MNGHA) enhances accuracy, efficiency, and patient outcomes [86]. Despite challenges including implementation costs, interoperability issues, and resistance to change [26], the integration of AI, telehealth, and wearable technologies [87], [88] represents a pivotal shift toward data-driven, patient-centric healthcare.

The parameters and macro-parameters identified in this section provide a structured view of healthcare system needs and priorities in Madinah. They present a cohesive picture of health in Madinah, shaped by the intersection of biological, behavioural, technological, and environmental factors. The results highlight key areas of focus, including surveillance, education, preventive interventions, and innovation, reflecting the multidimensional nature of healthcare development in the region. These collectively form the foundation for the subsequent analysis, where their alignment with AI governance frameworks is systematically evaluated using the PATHWAY framework. Table III provides a brief description of the five macro-parameters.

TABLE III. MACRO-PARAMETERS OF HEALTHCARE RESEARCH AND PRACTICE IN MADINAH

Macro-Parameter	Description
Infectious Diseases & Public Health Preparedness	Focuses on disease surveillance, outbreak response, epidemiological monitoring, and preparedness strategies for managing infectious threats at the population level.
Chronic Diseases & Lifestyle Health Issues	Addresses long-term health conditions such as diabetes and cardiovascular diseases, including lifestyle factors, prevention strategies, and population health management.
Paediatric & Maternal Health	Encompasses healthcare services, risk factors, and interventions related to maternal well-being, child development, and early-life health outcomes.
Environmental & Nutritional Health	Examines the impact of environmental conditions, dietary patterns, and nutrition-related factors on public health and disease prevalence.
Healthcare Systems & Technological Innovations	Covers healthcare delivery systems, digital health technologies, data-driven practices, and the integration of innovation into clinical and operational workflows.

V. AI GOVERNANCE ALIGNMENT WITH PUBLIC HEALTH PRIORITIES IN MADINAH

This section applies the PATHWAY framework to evaluate the alignment between the healthcare parameters identified in Section IV and existing AI governance frameworks. The analysis examines the extent to which these frameworks address key dimensions of healthcare research and practice in Madinah, with a focus on their coverage, relevance, and applicability. To provide a comprehensive perspective, the assessment is

conducted across both Saudi (Section IV A) and international governance frameworks (Section IV B), followed by a comparative synthesis that highlights key similarities, differences, and areas of convergence (Section IV C).

A. Saudi AI Governance Frameworks

The healthcare parameters identified in Section IV reflect a multidimensional health landscape in Madinah that intersects with Saudi Arabia's evolving AI and data governance ecosystem. Key national frameworks, including the Personal Data Protection Law [32], the Saudi AI Ethics Principles [12], the Generative Artificial Intelligence Guidelines for Government [33], the Human-Centered AI approach [34], the SFDA guidance on AI and ML-based medical devices [35], and the National Digital Government Strategy [36], collectively establish a foundation for data governance, ethical AI deployment, and digital transformation in healthcare contexts. These frameworks provide a structured basis for understanding how governance mechanisms relate to the healthcare priorities identified in Madinah.

Infectious Diseases & Public Health Preparedness illustrates the need for adaptive, data-driven protection measures. Variability in SARS-CoV-2 antibody responses among different demographic groups and individuals with comorbidities underscores the value of continuous monitoring and targeted vaccination strategies [62]. The Personal Data Protection Law [32] provides safeguards for such health surveillance by regulating lawful and minimal data processing, particularly when handling sensitive health data. The significant strain on ICU capacity due to delayed medical care, pre-existing illnesses, and high viral load at diagnosis [62], [63] reinforces the importance of reliable clinical analytics governed by the Saudi AI Ethics Principles, which emphasize fairness, transparency and human oversight. The unique risks associated with Hajj gatherings [19], [20] demonstrate the benefit of AI-assisted early warning systems in governmental settings, but these must comply with the Generative Artificial Intelligence Guidelines for Government that require accuracy, controlled use of restricted data and clarity when automated content informs public messaging. Hospital disaster readiness supported by accreditation and professional training [65] aligns with the National Digital Government Strategy through strengthened digital response infrastructure. Radiation exposure risk in orthopaedics [89] also points toward safe adoption of AI-assisted imaging tools, falling under SFDA guidance on AI and ML technologies-based medical devices to ensure validated clinical performance.

Chronic Diseases & Lifestyle Health Issues highlight the importance of preventive health technologies and patient empowerment. High consumption of sugary beverages and fast foods correlates with diabetes risk in the Madinah population including students [24], [68], and limited awareness of cardiovascular dangers affects early detection and lifestyle choices [23]. Human-Centered AI principles support digital tools that facilitate health education and behavioural change while maintaining inclusivity. Diabetic neuropathy results [90] indicate that remote monitoring and personalised care tools could assist disease management under ethical controls. Work stress and burnout among healthcare professionals [70] suggest that AI may help optimise workload and reduce fatigue when deployed responsibly under the Saudi AI Ethics Principles and accountability provisions. In cancer care, molecular determinants of colorectal cancer [69] signal the future integration of AI-enabled diagnostic systems whose safety would be regulated through SFDA guidance on AI and ML technologies-based medical devices. The influence of social media on cosmetic interventions [71] and persistent gaps in cervical cancer awareness [91] underline opportunities for reliable and culturally aligned generative AI public health communication under strict regulatory controls in the Generative Artificial Intelligence Guidelines for Government.

Paediatric & Maternal Health requires AI solutions that support equity and family wellbeing. High rates of childhood dental caries and unmet orthodontic needs [73], [78] stem from socio-economic and behavioural factors that can be targeted through digital outreach informed by the Saudi AI Ethics Principles [12] commitment to fairness. Reliance on natural remedies for oral conditions [92] highlights trust gaps that digital tools could help address if they align with cultural sensitivities emphasized in Human-Centered AI [34]. Limited awareness of paediatric eye diseases [93] suggests that parental engagement platforms could assist early detection when compliant with Personal Data Protection Law safeguards. Maternal health concerns, such as postpartum depression and breastfeeding challenges [75], show the need for tele-mental health and decision-support systems that enhance, rather than replace, professional support networks. Evidence associating caesarean delivery with childhood ASD risk [94] calls for decision aids that improve understanding of medical choices under strong human oversight and risk transparency. Widespread tobacco use and inadequate preparedness for dental trauma [95] similarly support the use of structured educational AI tools guided by the Generative Artificial Intelligence Guidelines for Government.

Environmental & Nutritional Health demonstrates how nutrition and environmental safety can both benefit from AI-supported monitoring. The protective effects of antioxidants in dates, particularly Ajwa varieties [96], support nutritional analytics that remain compliant with the Personal Data Protection Law [32] when integrated with health records. Meanwhile, heavy metal exposure at recycling sites poses significant respiratory risks [25], [82], and the Saudi AI Ethics Principles [12] and National Digital Government Strategy [36] could enable predictive environmental modelling that informs health protection, provided transparency and accountability are upheld. Nutritional strategies that reduce oxidative stress through metal chelation [83] further highlight the value of integrated analytics across food and environmental data.

Healthcare Systems & Technological Innovations illustrate direct intersections with national digital policy and biomedical AI regulation. Medication adherence research [97] suggests that predictive adherence analytics can support chronic disease control if aligned with Personal Data Protection Law principles of consent and security. The expanding digital health environment in Madinah's hospitals, including electronic records and telehealth [87], reinforces the National Digital Government Strategy objectives related to secure interoperability and efficient service delivery. As AI applications evolve within hospital workflows, they must

operate under the Saudi AI Ethics Principles and SFDA guidance on AI and ML technologies-based medical devices [35], where clinical impact is concerned.

Across these macro-parameters, Saudi AI governance frameworks demonstrate varying levels of alignment with the healthcare priorities identified in Madinah. Stronger alignment is observed in areas related to digital infrastructure, data governance, and healthcare system transformation, while more domain-specific applications are less explicitly represented within current governance provisions. Therefore, the analysis indicates that national frameworks provide a coherent foundation for supporting AI-enabled healthcare development, with alignment varying across different dimensions of the healthcare system.

B. International AI Governance Frameworks

The healthcare parameters identified in Section IV can be examined in relation to international AI governance frameworks, which provide structured approaches to regulating AI applications in healthcare and other high-impact domains. Prominent frameworks emphasise principles such as risk-based classification, transparency, accountability, human oversight, and data protection, offering a broader context for assessing how governance mechanisms align with the healthcare priorities observed in Madinah. National strategic frameworks such as the UK National AI Strategy [10], Japan's AI Strategy [46], Canada's AI Strategy for the Federal Public Service [48], and China's New Generation AIDP [11] articulate how these principles are operationalized at the policy level, particularly in guiding responsible AI adoption, public sector integration, and data governance across domains, including healthcare.

For instance, the challenges of managing infectious diseases, including COVID-19, and the unique public health risks during the Hajj pilgrimage resonate with the risk-based frameworks established by the EU AI Act [7]. Its classification of AI systems by risk underscores the importance of targeted oversight for applications such as pandemic surveillance, predictive analytics for ICU admissions, and diagnostic tools that can mitigate risks during large gatherings such as Hajj [62], [64]. Similarly, the US Blueprint for an AI Bill of Rights [8] and the Executive Order on AI [9] emphasize transparency, fairness, and accountability, principles that are directly relevant to building trust in AI-driven surveillance systems and diagnostic platforms in Madinah [67], [98].

The preparedness of hospitals in Madinah, as shown in studies on disaster readiness and ICU capacity, also intersects with the EU's Medical Device Regulation (MDR) [37] and In Vitro Diagnostic Regulation (IVDR) [38], which mandate ongoing monitoring and risk assessment for AI-enabled medical devices. These frameworks provide valuable guidance for ensuring that technologies adopted in Madinah's hospitals, including AI-driven diagnostic imaging or ICU monitoring, comply with safety and transparency requirements [99], [100]. Likewise, the US FDA's Software as a Medical Device (SaMD) framework [44] establishes robust standards for clinical evaluation, post-market surveillance, and patient safety, aligning with Madinah's emphasis on scalable healthcare infrastructure and continuous training of professionals [89], [101].

Chronic diseases and lifestyle health issues, such as diabetes, cardiovascular conditions, and cancer, represent a growing burden in Madinah. Here, AI governance frameworks addressing privacy and data protection play a pivotal role. The GDPR [39] in Europe stresses informed consent, transparency, and secure data management. These provisions are critical for managing sensitive health records, dietary surveys, and genetic information that underpin chronic disease research in Madinah [68], [102], [103]. Additionally, Japan's Act on the Protection of Personal Information (APPI) [47] and Canada's PIPEDA [49] reinforce the global consensus that privacy-preserving practices are indispensable when handling healthcare data for both research and clinical applications. This is particularly relevant for initiatives in Madinah aimed at improving public awareness of cardiovascular risks or managing diabetes through digital health tools [23].

Maternal and paediatric health results, including postpartum depression, congenital anomalies, and orthodontic needs, further highlight the need for ethical AI integration. The UNESCO Recommendation on the Ethics of AI [52] and the WHO guidance on AI in healthcare [28] call for equity, inclusiveness, and respect for human rights in deploying AI in clinical and community contexts. These frameworks support the design of AI-based early detection tools for congenital anomalies or mental health risks, ensuring that vulnerable groups, including women and children, benefit from technology without being exposed to additional ethical or privacy risks [73], [74], [75]. Similarly, UK GDPR [42] and the oversight of the Information Commissioner's Office (ICO) offer lessons in balancing innovation with protection of sensitive maternal and paediatric health data [93].

Environmental and nutritional health concerns in Madinah, such as antioxidant research on dates and heavy metal contamination at recycling sites, highlight another dimension of global AI governance. Regulations such as the EU NIS2 Directive [40] and Singapore's Regulatory Guidelines for Software Medical Devices [51] emphasize the role of cybersecurity and lifecycle monitoring in environmental and health-related AI systems. These frameworks can inform the design of AI-driven environmental monitoring tools that track pollution or food safety risks in Madinah, ensuring data integrity and reliability while protecting public health [96], [104]. Additionally, the EU's green deal [41] provides guidance on environmental issues by promoting sustainability, pollution reduction, and the integration of digital and data-driven approaches to support environmental monitoring and public health protection.

Finally, the adoption of digital health technologies and the validation of medication adherence tools in Madinah reflect the opportunities and challenges of technological innovation. The NIST AI RMF [45], the UK's use of regulatory sandbox [43], and China's Generative AI Regulation [50] collectively illustrate global recognition of the need for flexible but accountable oversight of emerging technologies. These initiatives provide a blueprint for Madinah's healthcare system to adopt AI responsibly, balancing innovation in digital health with safeguards for data security and patient trust [97], [105], [106].

International frameworks, such as those of the OECD [53] and IEEE [54], reinforce this balance by promoting transparency, robustness, and accountability across healthcare innovations [84], [107], [108].

Across these macro-parameters, international AI governance frameworks demonstrate a generally higher level of explicit coverage, particularly in areas involving high-risk applications, clinical decision support, and system-level accountability. This results in stronger alignment with several healthcare parameters identified in Madinah, especially those related to infectious disease management and healthcare system oversight. However, certain domains, including environmental and nutritional health, are less prominently represented within current governance provisions. Overall, international frameworks provide a comprehensive conceptual and regulatory basis, with alignment varying across different dimensions of healthcare.

C. Comparative Alignment

The comparative analysis of Saudi and international AI governance frameworks reveals both convergence and variation in their alignment with healthcare priorities identified in Madinah. Saudi frameworks demonstrate a strong emphasis on digital transformation, data governance, and the integration of AI within national development strategies. This results in relatively strong alignment with healthcare systems and technological innovation, as well as broader public health infrastructure.

In contrast, international frameworks provide more structured and mature approaches to AI governance, particularly through risk-based classifications, accountability mechanisms, and explicit consideration of high-risk applications in healthcare. These frameworks show stronger alignment with areas such as infectious disease monitoring, clinical decision support, and system-level oversight.

Differences also emerge in the extent to which specific healthcare domains are explicitly addressed. While some areas receive considerable attention across both governance contexts, others, including environmental and nutritional health, as well as context-specific aspects of paediatric and maternal healthcare, are less prominently reflected. These variations indicate differing priorities, levels of maturity, and approaches to operationalizing AI governance across contexts.

The analysis presented in this section suggests that Saudi frameworks are evolving and implementation-oriented, whereas international frameworks provide broader conceptual and regulatory structures. Together, they offer complementary perspectives that inform a more comprehensive understanding of AI governance in healthcare. Table IV summarises this comparative synthesis.

TABLE IV. COMPARATIVE ALIGNMENT OF AI GOVERNANCE FRAMEWORKS WITH MADINAH HEALTHCARE MACRO-PARAMETERS

Macro-Parameter	Saudi Frameworks	International Frameworks
Infectious Diseases & Public Health Preparedness	Moderately aligned through data governance and digital health initiatives; limited explicit AI-specific articulation	Strongly aligned through risk-based governance, monitoring, and system validation requirements
Chronic Diseases & Lifestyle Health Issues	Moderately reflected within preventive health and digital transformation strategies	Moderately aligned through clinical AI governance and decision-support frameworks
Paediatric & Maternal Health	Broadly addressed within general healthcare policies; limited explicit AI governance focus	Partially reflected under general healthcare AI risk classifications
Environmental & Nutritional Health	Indirectly reflected within sustainability and public health initiatives	Limited explicit representation within current AI governance frameworks
Healthcare Systems & Technological Innovations	Strongly aligned through national AI strategies and digital health transformation	Strongly aligned through comprehensive regulatory, ethical, and operational frameworks

VI. AI GOVERNANCE GAPS, RISK IMPLICATIONS, AND RECOMMENDATIONS

Despite strong alignment between Madinah's healthcare priorities and both Saudi Arabian and international AI governance frameworks, several important gaps remain where regulatory provisions do not fully accommodate the practical and contextual realities revealed by empirical results. Within the PATHWAY framework, this stage corresponds to the systematic identification of governance gaps and associated risks following alignment assessment, forming the basis for subsequent harmonisation through targeted recommendations. Foundational instruments such as the Personal Data Protection Law [32] and the Saudi AI Ethics Principles [12] provide robust safeguards for lawful data processing, oversight, and ethical AI deployment. Similarly, international risk-based frameworks, including the EU AI Act [7] and the US Blueprint for an AI Bill of Rights [8] emphasize accountability, transparency, and protection from harm. However, while these frameworks establish strong normative foundations, they provide limited operational guidance for population-specific, dynamic epidemiological surveillance.

Empirical evidence from Madinah demonstrates highly variable SARS-CoV-2 antibody responses and ICU demands influenced by comorbidities [62], [63], alongside intensified public health pressures during large-scale gatherings such as the Hajj [19], [20] . Although existing policies ensure responsible data governance, they do not sufficiently define structured pathways for continuous, event-driven epidemiological analytics tailored to demographic variability and localized health risks. This gap creates the risk of reduced responsiveness to rapidly evolving public health conditions. Addressing this limitation requires strengthening regulatory provisions to explicitly support context-sensitive, real-time surveillance systems that operate within clear accountability and privacy safeguards.

Similar regulatory limitations appear in the governance of behavioural and lifestyle-focused AI health tools. Human-Centered AI principles and privacy protections embedded within frameworks such as GDPR [39], PIPEDA [49], and APPI [47] reinforce secure handling of health data. Nevertheless, chronic disease drivers in Madinah, including dietary habits, cardiovascular risk unawareness, and lifestyle-related

conditions [102], [109], [110], reveal insufficient regulatory direction for AI systems designed to influence behavioural change, promote dietary awareness, or engage youth populations in preventive health practices. Ethical principles exist, yet they lack detailed guardrails for culturally sensitive behavioural intervention technologies. Without clearer governance mechanisms, such systems risk ethical ambiguity or misalignment with local values. Therefore, structured oversight frameworks for behavioural and lifestyle-oriented AI tools should be established to ensure culturally appropriate, prevention-focused digital health interventions that maintain strong privacy protections.

Further governance gaps emerge in maternal, paediatric, and mental health domains. International instruments such as UNESCO's Recommendation on the Ethics of AI [52] and WHO guidance on health AI [28] emphasize fairness, inclusivity, and protection of vulnerable populations, and Saudi frameworks similarly recognize societal and cultural values. However, Madinah results identify unmet orthodontic needs [111], limited awareness of paediatric eye diseases [93], postpartum mental health challenges [75], and broader maternal decision-support gaps. Existing regulatory structures lack specialised operational pathways for AI-assisted screening, tele-support systems, and child- or mother-focused digital health platforms. This creates ethical and trust-related risks for vulnerable groups. Developing specialised governance provisions tailored to maternal, paediatric, and mental health AI systems would strengthen ethical implementation, reinforce clinical oversight, and enhance trust by aligning digital interventions with family-centred and culturally grounded values.

Environmental and nutritional health governance also remains insufficiently integrated within current AI regulatory frameworks. The Saudi AI Ethics Principles [12] and the National Digital Government Strategy [36] promote responsible digital transformation, and instruments such as the EU NIS2 Directive [40] and Singapore guidelines [51] reinforce cybersecurity and lifecycle oversight for digital infrastructures that process cross-sector environmental and health data. However, empirical results from Madinah highlight health threats associated with heavy metal exposure at recycling sites [25], [82] and complex relationships between local dietary practices and antioxidant intake [112], [113]. Current governance structures do not clearly articulate how environmental monitoring data should be securely integrated with health analytics systems. This fragmentation limits transparent risk modelling and cross-sector accountability. Integrating environmental and nutritional data streams into AI governance architectures would enable secure cross-sector data sharing, strengthen predictive public health analytics, and enhance transparent risk management.

Oversight challenges also extend to AI-enabled medical devices and clinical workflow integration. Regulatory foundations are strong, with SFDA guidance [35], the EU MDR [37], IVDR [38], US SaMD requirements [44], and the Executive Order on AI [9] establishing safety, quality, and post-market monitoring standards. Nonetheless, Madinah results underscore the importance of scalable AI deployment in hospitals, workforce capacity-building, and digital systems that support treatment adherence and operational efficiency [26], [97], [101].

Current governance models prioritize technical compliance and safety validation but remain under-specified in relation to hospital-level adoption processes, workforce training, and real-world integration metrics. This creates implementation risks even where technical standards are satisfied. Expanding governance frameworks to include structured oversight for clinical adoption readiness, professional training, and workflow integration would strengthen real-world safety and performance.

Finally, governance of generative AI in public health communication remains underdeveloped. Although instruments such as the Generative Artificial Intelligence Guidelines for Government [33] promote transparency and responsible deployment, and broader international controls emphasize accountability, empirical results reveal persistent misinformation challenges in cosmetic decision-making [71], limited cervical cancer awareness [91], and inadequate responses to paediatric injuries [114]. Existing policies do not sufficiently ensure that AI-generated health content is culturally appropriate, clearly attributable, and responsive to local health literacy variations. Without strengthened safeguards, generative systems risk amplifying misinformation or eroding public trust. Strengthening governance for culturally aligned generative AI in healthcare communication would ensure transparent attribution, content validation mechanisms, and literacy-sensitive design standards that support improved community health outcomes.

Therefore, across these domains, a broader structural pattern emerges. Many governance frameworks articulate high-level principles of fairness, accountability, privacy, and transparency, yet lack detailed operational pathways tailored to specific demographic, epidemiological, environmental, and institutional contexts. This operationalization gap, combined with limited cross-sector integration and insufficient implementation-readiness mechanisms, constrains the full realization of AI-enabled healthcare benefits in Madinah. From a PATHWAY perspective, these gaps highlight the need for structured harmonisation through targeted recommendations and governance refinement. Addressing these interconnected gaps through targeted regulatory refinement will enable governance systems not only to uphold foundational ethical standards but also to support culturally aligned, context-sensitive, and operationally scalable AI deployment. In doing so, Madinah's healthcare ecosystem can advance equitable, resilient, and technology-enabled outcomes aligned with both national priorities and international best practices.

Fig. 5 synthesises the identified AI governance gaps in Madinah's healthcare ecosystem, highlighting their associated risks and the corresponding recommendations required to support context-sensitive, scalable, and ethically aligned AI deployment. Moreover, Table V provides a structured overview of domain-specific AI governance gaps in Madinah's public health system, linking each gap to its corresponding risk implications and targeted governance recommendations.

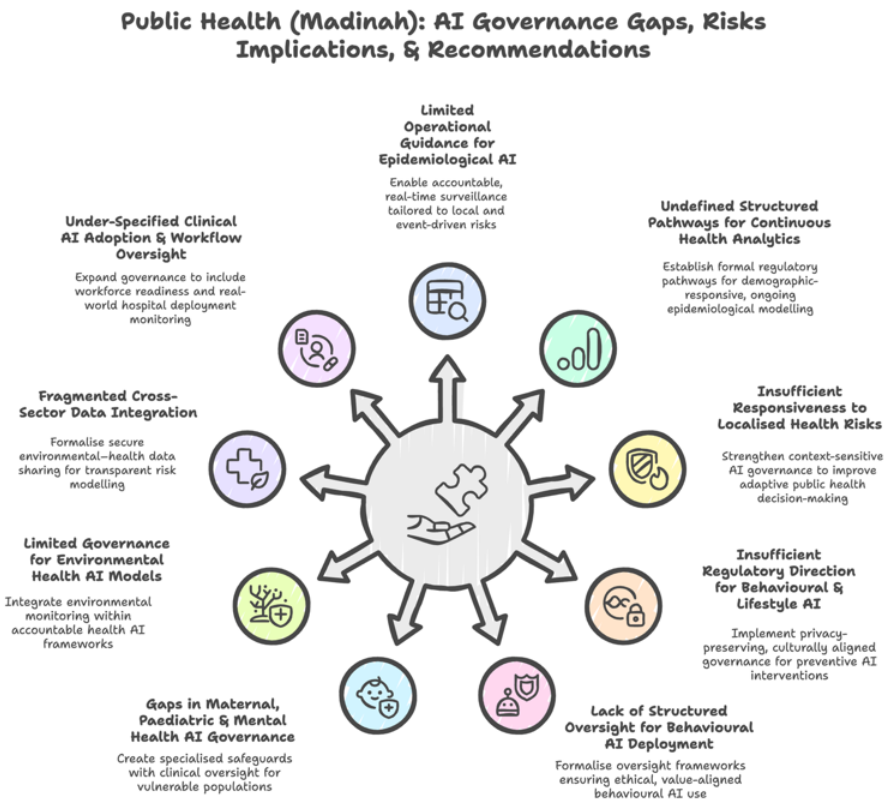


Fig. 5. Public health (Madinah): AI governance gaps, risk implications, and recommendations.

TABLE V. DOMAIN-LEVEL AI GOVERNANCE GAPS, RISK IMPLICATIONS, AND RECOMMENDED ACTIONS IN PUBLIC HEALTH (MADINAH)

Domain	Key Governance Gap	Risk Implication	Recommendation
Infectious Diseases & Public Health Preparedness	Lack of explicit governance for real-time epidemiological AI systems	Reduced responsiveness to dynamic outbreaks	Introduce regulatory pathways for real-time surveillance and model validation
Chronic Diseases & Lifestyle Health Issues	Limited guidance for behavioural and preventive AI systems	Ethical ambiguity and misalignment with cultural values	Develop governance for culturally sensitive behavioural AI tools
Paediatric & Maternal Health	Absence of specialised provisions for vulnerable populations	Safety and trust risks in sensitive healthcare domains	Establish tailored governance for child and maternal AI applications
Environmental & Nutritional Health	Weak integration of environmental data in AI governance	Fragmented risk modelling and limited cross-sector insight	Enable secure integration of environmental and health data systems
Healthcare Systems & Technological Innovation	Insufficient focus on implementation and workforce readiness	Deployment and adoption risks in real-world settings	Expand governance to include training, workflow integration, and monitoring
Generative AI in Public Health	Limited safeguards for health communication systems	Risk of misinformation and reduced public trust	Introduce governance for culturally aligned and validated AI-generated content
Mass Gathering and Event-driven Public Health Systems	Lack of governance for AI systems operating in large-scale, dynamic population environments (e.g., Hajj)	Reduced responsiveness and coordination in high-density, rapidly evolving public health scenarios	Develop governance frameworks for event-driven, real-time population health analytics and coordination systems
Demographic Variability and Personalised Public Health	Limited governance for AI models handling heterogeneous population characteristics and responses	Reduced fairness, accuracy, and reliability in population-level predictions and interventions	Introduce governance provisions for adaptive, demographic-aware AI systems and model validation across diverse groups
Clinical Workflow Integration and Adoption Readiness	Governance frameworks do not sufficiently address hospital-level integration, workflow embedding, and adoption processes	Implementation failures despite technical compliance; reduced efficiency and usability in clinical settings	Extend governance to include operational readiness, workflow integration standards, and clinical deployment guidelines
Real-world Monitoring and Lifecycle Oversight	Insufficient operationalization of continuous monitoring, validation, and feedback mechanisms for deployed AI systems	Undetected model drift, reduced reliability, and long-term performance degradation	Establish lifecycle governance models with continuous evaluation, auditing, and post-deployment monitoring
AI-enabled Public Health Communication Systems	Limited governance for AI-mediated health communication beyond general principles (including but not limited to generative AI)	Risk of misinformation, low health literacy alignment, and erosion of public trust	Develop governance mechanisms for transparent, culturally aligned, and validated AI-driven health communication systems

VII. Conclusion

This study has examined the alignment between healthcare system needs and AI governance frameworks through an evidence-driven and context-sensitive approach. By integrating large-scale empirical analysis of healthcare research and practice in Madinah with structured governance evaluation, the study demonstrates how governance frameworks can be systematically assessed against real-world system requirements. The combined application of the PEARL and PATHWAY frameworks enables a coherent analytical pipeline that connects empirical evidence, potential AI applications, governance alignment, and the identification of gaps, risks, and recommendations.

The results indicate that while existing Saudi Arabian and international AI governance frameworks provide strong foundations in terms of ethical principles, data protection, and risk-based regulatory models, they remain insufficiently operationalized for domain-specific and context-sensitive healthcare applications. In particular, gaps emerge in areas requiring continuous and adaptive responses, including epidemiological surveillance, behavioural and lifestyle health interventions, maternal and paediatric care, environmental health integration, clinical workflow implementation, and the governance of AI-generated public health information. These limitations reflect a broader structural issue in which high-level governance, although necessary, does not fully translate into operational and context-aware mechanisms within complex healthcare environments.

This work contributes to addressing this gap by demonstrating the value of linking empirical healthcare evidence with governance analysis. Rather than relying solely on abstract regulatory principles or isolated case studies, the approach adopted in this study enables a systematic and data-informed understanding of how governance frameworks interact with real-world healthcare needs. In doing so, it highlights the importance of moving towards governance models that are not only principle-based but also supported by structured evidence, capable of capturing evolving system dynamics, and responsive to local conditions.

Furthermore, our study underscores the importance of governance approaches that are supported by continuous and iterative analysis of healthcare data, research developments, and societal needs, forming a structured feedback loop in which evidence generation, governance evaluation, and policy refinement are perpetually linked (see our earlier works, e.g., [115], [116], [117]). Such an approach enables governance systems to evolve in response to emerging risks, new evidence, and changing public health conditions, while maintaining transparency, traceability, and accountability. By embedding this continuous evaluation and refinement cycle into governance processes, policymakers can develop more responsive and adaptive strategies, improve the reliability of health communication, and strengthen public trust by ensuring that AI-enabled systems remain aligned with both scientific evidence and community realities. This perspective is consistent with emerging approaches that emphasise systematic, data-driven analysis as a foundation for ongoing policy monitoring and refinement.

The study also positions Madinah as a meaningful empirical environment for governance evaluation, demonstrating how localised analysis can reveal insights that are not visible in global or purely conceptual studies. The methodological approach and results are therefore not limited to this context but offer a generalizable framework for analysing AI governance in other regions and domains characterised by complex, dynamic, and data-intensive systems.

Therefore, advancing AI governance in healthcare requires moving beyond high-level principles towards evidence-driven, domain-informed, and context-aware operationalization. By bridging empirical analysis with governance evaluation, this study contributes to the development of more adaptive, scalable, trustworthy, and sustainable governance strategies, supporting the effective and responsible integration of AI into healthcare systems.

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] A. Al Kuwaiti et al., "A Review of the Role of Artificial Intelligence in Healthcare," J. Pers. Med., vol. 13, no. 6, Jun. 2023, doi: 10.3390/jpm13060951.

[2] X. Gao, P. He, Y. Zhou, and X. Qin, "Artificial Intelligence Applications in Smart Healthcare: A Survey," Future Internet 2024, Vol. 16, vol. 16, no. 9, Aug. 2024, doi: 10.3390/fi16090308.

[3] R. Alsaigh et al., "Harmonizing AI governance regulations and neuroinformatics: perspectives on privacy and data sharing," Front. Neuroinform., vol. 18, Dec. 2024, doi: 10.3389/fninf.2024.1472653.

[4] R. Alharthi, R. Mehmood, and A. Albeshri, "A Systematic, Scalable, and Interpretable Mapping of Artificial Intelligence Research in Leukemia Using a Hybrid Machine Learning and Qualitative Framework," Electronics 2026, Vol. 15, vol. 15, no. 5, Mar. 2026, doi: 10.3390/electronics15051078.

[5] 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.

[6] S. Alghamdi, R. Mehmood, F. Alqurashi, and A. Alzahrani, "Paving the Roadmap for XAI and IML in Healthcare: Data-Driven Discoveries and the FIXAIH Framework," Nov. 2024, doi: 10.2139/SSRN.5086713.

[7] European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 an."

[8] The White House, "Blueprint for an AI Bill of Rights - MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE," 2022.

[9] CISA, "Executive Order on Improving the Nation's Cybersecurity." Accessed: Oct. 30, 2025. [Online]. Available: https://www.cisa.gov/topics/cybersecurity-best-practices/executive-order-improving-nations-cybersecurity

[10] UK Government, "National AI Strategy." Accessed: Nov. 06, 2024. [Online]. Available: https://www.gov.uk/government/publications/national-ai-strategy

[11] G. Webster, R. Creemers, E. Kania, and P. Triolo, "Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)," DigiChina, Standford University.

[12] Saudi Data and AI Authority (SDAIA), "AI Ethics Principles," 2023, Accessed: Nov. 17, 2025. [Online]. Available: https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf

[13] C. Bublitz, F. Molnár-Gábor, and S. R. Soekadar, "Implications of the novel EU AI Act for neurotechnologies," Neuron, vol. 112, no. 18, pp. 3013–3016, Sep. 2024, doi: 10.1016/J.NEURON.2024.08.011.

[14] M. A. Yar, M. Hamdan, M. Anshari, N. L. Fitriyani, and M. Syafrudin, "Governing with Intelligence: The Impact of Artificial Intelligence on Policy Development," Information 2024, Vol. 15, Page 556, vol. 15, no. 9, p. 556, Sep. 2024, doi: 10.3390/INFO15090556.

[15] I. G. Cohen, T. Evgeniou, S. Gerke, and T. Minssen, "The European artificial intelligence strategy: implications and challenges for digital health," Lancet Digit. Health, vol. 2, no. 7, pp. e376–e379, Jul. 2020, doi: 10.1016/S2589-7500(20)30112-6.

[16] T. Minssen, E. Vayena, and I. G. Cohen, "The Challenges for Regulating Medical Use of ChatGPT and Other Large Language Models," JAMA, vol. 330, no. 4, pp. 315–316, Jul. 2023, doi: 10.1001/JAMA.2023.9651.

[17] B. C. Stahl and T. Leach, "Assessing the ethical and social concerns of artificial intelligence in neuroinformatics research: an empirical test of the European Union Assessment List for Trustworthy AI (ALTAI)," AI and Ethics, vol. 3, no. 3, pp. 745–767, Aug. 2023, doi: 10.1007/S43681-022-00201-4.

[18] L. Mulugeta et al., "Credibility, replicability, and reproducibility in simulation for biomedicine and clinical applications in neuroscience," Front. Neuroinform., vol. 12, p. 359627, Apr. 2018, doi: 10.3389/FNINF.2018.00018/BIBTEX.

[19] N. M. Abdalla, A. M. Abdelgani, A. A. Osman, and M. N. Mohamed, "Demographical and population dynamics impact on public health of cutaneous Leishmaniasis in Al-Madinah Almonawra, Saudi Arabia," Afr. Health Sci., vol. 19, no. 3, p. 2421, Sep. 2019, doi: 10.4314/AHS.V19I3.16.

[20] F. Sharaf et al., "Ethnicity is a major determinant of the pattern of dermatological diseases among pilgrims during the Hajj in 2019," Dermatol. Reports, vol. 13, no. 1, Mar. 2021, doi: 10.4081/DR.2021.8934.

[21] General Authority for Statistics (GASTAT), "Umrah Statistics." Accessed: Mar. 24, 2026. [Online]. Available: https://www.stats.gov.sa/en/statistics-tabs?tab=436312&category=124334

[22] D. V. Parums, "Editorial: Infectious Disease Surveillance Using Artificial Intelligence (AI) and its Role in Epidemic and Pandemic Preparedness," Med. Sci. Monit., vol. 29, pp. e941209-1, 2023, doi: 10.12659/MSM.941209.

[23] M. M. Alluqmani, N. R. Almshhen, R. A. Alotaibi, O. Y. Aljardi, and H. M. Zahid, "Public Awareness of Ischemic Stroke in Medina city, Kingdom of Saudi Arabia," Neurosciences, vol. 26, no. 2, p. 134, 2021, doi: 10.17712/NSJ.2021.2.20200105.

[24] W. A. Mumena et al., "Knowledge, attitudes, and practices toward added sugar consumption among female undergraduate students in Madinah, Saudi Arabia: A cross-sectional study," Nutrition, vol. 79–80, p. 110936, Nov. 2020, doi: 10.1016/J.NUT.2020.110936.

[25] M. Tasleem, A. A. A. A. El-Sayed, W. M. Hussein, and A. Alrehaily, "Pseudomonas putida Metallothionein: Structural Analysis and Implications of Sustainable Heavy Metal Detoxification in Madinah," Toxics, vol. 11, no. 10, Oct. 2023, doi: 10.3390/TOXICS11100864.

[26] H. F. Al Otaybi, R. M. Al-Raddadi, and F. H. Bakhamees, "Performance, Barriers, and Satisfaction of Healthcare Workers Toward Electronic Medical Records in Saudi Arabia: A National Multicenter Study," Cureus, vol. 14, no. 2, Feb. 2022, doi: 10.7759/CUREUS.21899.

[27] P. Ahadian, W. Xu, D. Liu, and Q. Guan, "Ethics of trustworthy AI in healthcare: Challenges, principles, and practical pathways," Neurocomputing, vol. 661, p. 131942, Jan. 2026, doi: 10.1016/j.neucom.2025.131942.

[28] WHO, "Ethics and governance of artificial intelligence for health: WHO guidance Executive summary," World Health Organization. Accessed: Nov. 17, 2024. [Online]. Available: https://www.who.int/publications/i/item/9789240037403

[29] J. Morley, L. Murphy, A. Mishra, I. Joshi, and K. Karpathakis, "Governing Data and Artificial Intelligence for Health Care: Developing an International Understanding," JMIR Form. Res., vol. 6, no. 1, p. e31623, Jan. 2022, doi: 10.2196/31623.

[30] N. Monteiro and V. Singh, "The wheel of artificial intelligence governance," Sustainable Futures, vol. 10, p. 101279, Dec. 2025, doi: 10.1016/j.sftr.2025.101279.

[31] R. Alsaigh, R. Mehmood, I. Katib, and T. Yigitcanlar, "Governing AI in Society: Explainable Analysis of Research using the PEARL Methodology, The Frame-AI Framework, and Regulatory Alignment Gaps," SSRN Electronic Journal, Dec. 2025, doi: 10.2139/SSRN.6009614.

[32] Saudi Data and AI Authority (SDAIA), "Personal Data Protection Law." Accessed: Nov. 19, 2024. [Online]. Available: https://sdaia.gov.sa/en/SDAIA/about/Documents/Personal Data English V2-23April2023- Reviewed-.pdf

[33] Saudi Data and AI Authority (SDAIA), "Generative Artificial Intelligence for Government," 2025, Accessed: Nov. 17, 2025. [Online]. Available: https://sdaia.gov.sa/en/SDAIA/about/Files/GenAIGuidelinesForGovernmentENCompressed.pdf

[34] Digital Government Authority, "Human-Centered AI." Accessed: Nov. 18, 2025. [Online]. Available: https://dga.gov.sa/en/digital-knowledge/human-centered-AI

[35] Saudi Food and Drug Authority, "Guidance on Artificial Intelligence (AI) and Machine Learning (ML) technologies based Medical Devices." Accessed: Nov. 19, 2024. [Online]. Available: https://www.sfda.gov.sa/sites/default/files/2023-01/MDS-G010ML.pdf

[36] National Platform (National Portal), "Digital Government Strategy." Accessed: Nov. 18, 2025. [Online]. Available: https://my.gov.sa/en/content/digital-strategy#section-6

[37] European Commission, "The European Union Medical Device Regulation – Regulation (EU) 2017/745 (EU MDR)," The European Union Medical Device Regulation. Accessed: Nov. 24, 2024. [Online]. Available: https://eumdr.com/

[38] European Commission, "The European Union In Vitro Diagnostics Regulation – Regulation (EU) 2017/746 (EU IVDR)." Accessed: Nov. 06, 2024. [Online]. Available: https://euivdr.com/

[39] European Union, "General Data Protection Regulation (GDPR) – Legal Text." Accessed: Nov. 08, 2024. [Online]. Available: https://gdpr-info.eu/

[40] European Commission, "NIS2 Directive: securing network and information systems-Shaping Europe's digital future." Accessed: Nov. 26, 2025. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/nis2-directive

[41] European Commission, "The European Green Deal." Accessed: Nov. 26, 2025. [Online]. Available: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

[42] ICO, "The UK GDPR." Accessed: Nov. 06, 2024. [Online]. Available: https://ico.org.uk/for-organisations/data-protection-and-the-eu/data-protection-and-the-eu-in-detail/the-uk-gdpr/

[43] ICO, "Regulatory Sandbox." Accessed: Nov. 13, 2024. [Online]. Available: https://ico.org.uk/for-organisations/advice-and-services/regulatory-sandbox/

[44] FDA, "Software as a Medical Device (SaMD)." Accessed: Nov. 13, 2024. [Online]. Available: https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd

[45] NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," 2023. doi: 10.6028/NIST.AI.100-1.

[46] Government of Japan, "AI Strategy 2022," 2022.

[47] Japanese Law Translation, "Act on the Protection of Personal Information - English -." Accessed: Nov. 13, 2024. [Online]. Available: https://www.japaneselawtranslation.go.jp/en/laws/view/4241/en

[48] Government of Canada, "AI Strategy for the Federal Public Service 2025-2027: Overview." Accessed: Mar. 25, 2026. [Online]. Available: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/gc-ai-strategy-overview.html

[49] Office of the Privacy Commissioner of Canada, "The Personal Information Protection and Electronic Documents Act (PIPEDA)." Accessed: Nov. 13, 2024. [Online]. Available:

https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/

[50] Seaton Huang, Helen Toner, Zac Haluza, Rogier Creemers, and Graham Webster, "Translation: Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment." Accessed: Oct. 21, 2025. [Online]. Available: https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/

[51] Health Sciences Authority, "Regulatory Guidelines for Software Medical Devices-A Life Cycle Approach Revision 2.0," 2022. Accessed: Nov. 19, 2024. [Online]. Available: https://www.hsa.gov.sg/docs/default-source/hprg-mdb/guidance-documents-for-medical-devices/regulatory-guidelines-for-software-medical-devices---a-life-cycle-approach_r2-(2022-apr)-pub.pdf

[52] UNESCO, "Recommendation on The Ethics of Artificial Intelligence," 2022. Accessed: Nov. 17, 2024. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000381137

[53] OECD, "AI principles." Accessed: Oct. 31, 2025. [Online]. Available: https://www.oecd.org/en/topics/sub-issues/ai-principles.html

[54] IEEE SA, "The IEEE Global Initiative 2.0 on Ethics of Autonomous and Intelligent Systems." Accessed: Jan. 13, 2024. [Online]. Available: https://standards.ieee.org/industry-connections/activities/ieee-global-initiative/

[55] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.

[56] UK Government, "Crafting an intended purpose in the context of Software as a Medical Device (SaMD)." Accessed: Nov. 19, 2024. [Online]. Available: https://www.gov.uk/government/publications/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd

[57] UK Government, "Good Machine Learning Practice for Medical Device Development: Guiding Principles." Accessed: Nov. 19, 2024. [Online]. Available: https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles/good-machine-learning-practice-for-medical-device-development-guiding-principles

[58] European Commission, "Clinical trials - Regulation EU No 536/2014." Accessed: Nov. 06, 2024. [Online]. Available: https://health.ec.europa.eu/medicinal-products/clinical-trials/clinical-trials-regulation-eu-no-5362014_en

[59] R. Hussein et al., "Advancing healthcare AI governance through a comprehensive maturity model based on systematic review," npj Digital Medicine 2026 9:1, vol. 9, no. 1, pp. 236-, Feb. 2026, doi: 10.1038/s41746-026-02418-7.

[60] S. Reddy, S. Allan, S. Coghlan, and P. Cooper, "A governance model for the application of AI in health care," Journal of the American Medical Informatics Association, vol. 27, no. 3, pp. 491–497, Mar. 2020, doi: 10.1093/jamia/ocz192.

[61] A. Aqeeli, A. T. Alsharif, M. Tennant, and E. Kruger, "School Dental Services Theoretical Model-Based on Geographic Information System in Al-Madinah, Saudi Arabia," Children (Basel), vol. 10, no. 2, Feb. 2023, doi: 10.3390/CHILDREN10020186.

[62] S. A. Amer et al., "Role of neutrophil to lymphocyte ratio as a prognostic indicator for COVID-19," Health Sci. Rep., vol. 4, no. 4, Dec. 2021, doi: 10.1002/HSR2.442.

[63] S. T. Alhoufie et al., "Epidemiological Characteristics and Outcomes Predictors for Intensive Care Unit COVID-19 Patients in Al-Madinah, Saudi Arabia. Retrospective Cohort Study," Infect. Drug Resist., vol. 16, pp. 5573–5586, 2023, doi: 10.2147/IDR.S419724.

[64] V. T. Hoang et al., "Environmental investigation of respiratory pathogens during the Hajj 2016 and 2018," Travel Med. Infect. Dis., vol. 33, p. 101500, Jan. 2020, doi: 10.1016/J.TMAID.2019.101500.

[65] O. G. Baker, A. A. Alamri, and A. E. Aboshaiqah, "A descriptive study to analyse the disaster preparedness among Saudi nurses through self-regulation survey," J. Nurs. Manag., vol. 27, no. 7, pp. 1479–1484, Oct. 2019, doi: 10.1111/JONM.12833.

[66] A. Al-Zalabani, O. A. AlThobyane, A. H. Alshehri, A. O. Alrehaili, M. O. Namankani, and O. H. Aljafri, "Prevalence of Klebsiella pneumoniae Antibiotic Resistance in Medina, Saudi Arabia, 2014-2018," Cureus, vol. 12, no. 8, Aug. 2020, doi: 10.7759/CUREUS.9714.

[67] A. M. Almotairy, W. A. Sheikh, A. A. A. Joraid, A. A. Bajwi, M. S. F. Alharbi, and S. A. R. Al-Dubai, "Association between knowledge of influenza vaccine and vaccination status among general population attending primary health care centers in Al-Madinah, Saudi Arabia," J. Family Med. Prim. Care, vol. 8, no. 9, p. 2971, 2019, doi: 10.4103/JFMPC.JFMPC_547_19.

[68] K. Basham et al., "Public awareness of early symptoms of acute myocardial infarction among Saudi population," J. Family Med. Prim. Care, vol. 10, no. 10, p. 3785, 2021, doi: 10.4103/JFMPC.JFMPC_449_21.

[69] N. Mulla, A. Alshareef, A. R. Syed, and M. Al-Jahel, "Clinico-Pathological Study of K-ras Mutations in Colorectal Tumors: A Single-Center Retrospective Study of 51 Patients in Madinah, Saudi Arabia," Cureus, vol. 12, no. 8, Aug. 2020, doi: 10.7759/CUREUS.9978.

[70] S. A. R. Aldubai, A. M. Aljohani, A. G. Alghamdi, K. S. Alghamdi, K. Ganasegeran, and A. M. Yenbaawi, "Prevalence and associated factors of burnout among family medicine residents in Al Madina, Saudi Arabia," J. Family Med. Prim. Care, vol. 8, no. 2, p. 657, 2019, doi: 10.4103/JFMPC.JFMPC_268_18.

[71] E. E. Sindi et al., "Factors That Affect the Likelihood of Undergoing Cosmetic Procedures Among the General Population of Western Region, Saudi Arabia: A Cross-Sectional Study," Cureus, vol. 15, no. 12, Dec. 2023, doi: 10.7759/CUREUS.49792.

[72] N. H. Almohammadi, "The pattern of gynecological malignancies in Al-Madinah Al-Munawarah region, Saudi Arabia: An overview of 6 years," Saudi Med. J., vol. 43, no. 3, pp. 283–290, Mar. 2022, doi: 10.15537/SMJ.2022.43.3.20210888.

[73] R. T. Ashky, N. M. Althagafi, B. H. Alsaati, R. A. Alharbi, S. A. Kassim, and A. T. Alsharif, "Self-Perception Of Malocclusion And Barriers To Orthodontic Care: A Cross-Sectional Study In Al-Madinah, Saudi Arabia," Patient Prefer. Adherence, vol. 13, p. 1723, 2019, doi: 10.2147/PPA.S219564.

[74] H. J. Sabbagh et al., "Incidence of non-syndromic orofacial cleft during the COVID-19 pandemic in Saudi Arabia," BMC Public Health, vol. 23, no. 1, pp. 1–11, Dec. 2023, doi: 10.1186/S12889-023-17270-7/TABLES/6.

[75] R. Abdulqader Alrehaili and R. Albelowi, "The Prevalence of Postpartum Depression and the Related Risk Factors in Primary Health Care, Al-Madinah, Saudi Arabia," Cureus, vol. 14, no. 2, Feb. 2022, doi: 10.7759/CUREUS.22681.

[76] O. Suliman, A. M. Alraddadi, F. M. Almutairi, F. A. Almohammadi, and Z. A. Albakri, "Age at Initial Cleft Lip Repair Among Children in Al Madinah Region," Cureus, vol. 15, no. 11, Nov. 2023, doi: 10.7759/CUREUS.49089.

[77] A. Murad, M. J. Renfrew, A. Symon, and H. Whitford, "Understanding factors affecting breastfeeding practices in one city in the Kingdom of Saudi Arabia: an interpretative phenomenological study," Int. Breastfeed. J., vol. 16, no. 1, pp. 1–9, Dec. 2021, doi: 10.1186/S13006-020-00350-4/TABLES/2.

[78] S. Kassim, H. Bakeer, S. Alghazy, Y. Almaghraby, W. Sabbah, and A. Alsharif, "Socio-Demographic Variation, Perceived Oral Impairment and Oral Impact on Daily Performance among Children in Saudi Arabia," Int. J. Environ. Res. Public Health, vol. 16, no. 14, Jul. 2019, doi: 10.3390/IJERPH16142450.

[79] S. Alqadi et al., "Perception and Attitude of Parents of Children With Orofacial Clefts Regarding the Use of Presurgical Orthopedics and Feeding Obturators," Cureus, vol. 15, no. 9, Sep. 2023, doi: 10.7759/CUREUS.46131.

[80] S. Khalid, A. Ahmad, and M. Kaleem, "Antioxidant activity and phenolic contents of Ajwa date and their effect on lipo-protein profile," Functional Foods in Health and Disease, vol. 7, no. 6, pp. 396–410, Jun. 2017, doi: 10.31989/FFHD.V7I6.337.

[81] P. K. Vayalil, "Date fruits (Phoenix dactylifera Linn): an emerging medicinal food," Crit. Rev. Food Sci. Nutr., vol. 52, no. 3, pp. 249–271, 2012, doi: 10.1080/10408398.2010.499824.

[82] M. A. Alnuwaiser, "Evaluation of Heavy Metals in Soil Wastewater Stream," Int. J. Anal. Chem., vol. 2022, 2022, doi: 10.1155/2022/2522840.

[83] I. Romieu, F. Castro-Giner, N. Kunzli, and J. Sunyer, "Air pollution, oxidative stress and dietary supplementation: a review," European Respiratory Journal, vol. 31, no. 1, pp. 179–197, Dec. 2007, doi: 10.1183/09031936.00128106.

[84] A. A. Naqvi et al., "Translation and validation of the Arabic version of the General Medication Adherence Scale (GMAS) in Saudi patients with chronic illnesses," Saudi Pharm. J., vol. 28, no. 9, pp. 1055–1061, Sep. 2020, doi: 10.1016/J.JSPS.2020.07.005.

[85] R. Aljohani, Z. Aljohani, R. Aljohani, and R. Alsaidalani, "Saudi cultural adaptation of the 'compliance questionnaire of Rheumatology' for Rheumatoid arthritis patients on disease modifying anti-rheumatic drugs (DMARDs)," Saudi Pharm. J., vol. 29, no. 5, pp. 377–383, May 2021, doi: 10.1016/J.JSPS.2021.03.007.

[86] C. C. Yousef et al., "Adoption of a Personal Health Record in the Digital Age: Cross-Sectional Study," J. Med. Internet Res., vol. 22, no. 10, Oct. 2020, doi: 10.2196/22913.

[87] M. S. Alauddin, A. S. Baharuddin, and M. I. M. Ghazali, "The Modern and Digital Transformation of Oral Health Care: A Mini Review," Healthcare 2021, Vol. 9, Page 118, vol. 9, no. 2, p. 118, Jan. 2021, doi: 10.3390/HEALTHCARE9020118.

[88] A. A. Almohanna, K. T. Win, and S. Meedya, "Effectiveness of internet-based electronic technology interventions on breastfeeding outcomes: Systematic review," J. Med. Internet Res., vol. 22, no. 5, p. e17361, May 2020, doi: 10.2196/17361.

[89] Y. A. Alshabi et al., "Knowledge Regarding Ionizing Radiation Exposure Safety Among Orthopedic Surgeons at Hospitals in Al-Madinah," Cureus, vol. 14, no. 10, Oct. 2022, doi: 10.7759/CUREUS.30738.

[90] R. A. Sendi, A. M. Mahrus, R. M. Saeed, M. A. Mohammed, and S. A. R. Al-Dubai, "Diabetic peripheral neuropathy among Saudi diabetic patients: A multicenter cross-sectional study at primary health care setting," J. Family Med. Prim. Care, vol. 9, no. 1, p. 197, 2020, doi: 10.4103/JFMPC.JFMPC_927_19.

[91] H. M. Yousif, A. M. Albasri, M. M. Alshanqite, and H. M. Missawi, "Histopathological Patterns and Characteristics of Abnormal Cervical Smear in Madinah Region of Saudi Arabia," Asian Pac. J. Cancer Prev., vol. 20, no. 5, p. 1303, 2019, doi: 10.31557/APJCP.2019.20.5.1303.

[92] N. S. Dar-Odeh, R. A. Gasim, S. M. Binsaad, S. Abu-Hammad, and O. A. Abu-Hammad, "Use of natural remedies to treat oral diseases among female patients in Al Madinah, western Saudi Arabia," J. Complement. Integr. Med., vol. 16, no. 3, 2019, doi: 10.1515/JCIM-2018-0032.

[93] A. M. Surrati et al., "Parents' Awareness and Perception of Children's Eye Diseases in Madinah, Saudi Arabia: A Cross-Sectional Study," Cureus, vol. 14, no. 2, Feb. 2022, doi: 10.7759/CUREUS.22604.

[94] A. H. Al-Zalabani, A. H. Al-Jabree, and Z. A. Zeidan, "Is cesarean section delivery associated with autism spectrum disorder?," Neurosciences (Riyadh), vol. 24, no. 1, pp. 11–15, 2019, doi: 10.17712/NSJ.2019.1.20180303.

[95] Z. F. Alharbi and M. A. Habibullah, "Knowledge, Attitudes, and Practices of School Teachers Regarding Dental Trauma and its Emergency Management in Madinah, Saudi Arabia: A Questionnaire-Based Online Cross-Sectional Survey," J. Pharm. Bioallied Sci., vol. 15, no. Suppl 1, pp. S775–S782, Jul. 2023, doi: 10.4103/JPBS.JPBS_107_22.

[96] S. M. A. Hassan et al., "The Preventive and Therapeutic Effects of Ajwa Date Fruit Extract Against Acute Diclofenac Toxicity-Induced Colopathy: An Experimental Study," Drug Des. Devel. Ther., vol. 16, p. 2601, 2022, doi: 10.2147/DDDT.S344247.

[97] R. M. Alofi, L. S. Alrohaily, R. A. Jan, S. L. Alsaedi, F. A. Mahrous, and M. M. Alreefi, "Adherence to Levothyroxine Treatment Among Patients With Hypothyroidism in Madinah, Saudi Arabia: A Cross-Sectional Study," Cureus, vol. 15, no. 6, Jun. 2023, doi: 10.7759/CUREUS.40686.

[98] W. H. Mahallawi et al., "Assessment of SARS-CoV-2 Anti-Spike IgG Antibody in Women and Children in Madinah, Saudi Arabia: A Single-Center Study," Int. J. Environ. Res. Public Health, vol. 18, no. 19, Oct. 2021, doi: 10.3390/IJERPH18199971.

[99] Z. Z. Al-Ahmadey, A. M. Aljohani, S. S. Al-Ahmadi, A. Alruwetei, and R. A. Alharbi, "Clinical Characteristics among Patients with COVID-19: A Single-Center Experience from Medina, Saudi Arabia," J. Pure Appl. Microbiol., vol. 16, no. 3, pp. 1673–1681, Sep. 2022, doi: 10.22207/JPAM.16.3.06.

[100] A. Alsaedi, S. Sukeri, N. M. Yaccob, and A. Almazroea, "Impact of the Accreditation Program of the Saudi Central Board for Accreditation of Healthcare Institutions on the Safety Dimension of the Institute of Medicine Quality," 2023, doi: 10.2147/JMDH.S410925.

[101] F. A. Alrehily et al., "Establishing local diagnostic reference levels for computed tomography examinations using size-specific dose estimates," Saudi Med. J., vol. 44, no. 8, pp. 761–766, Aug. 2023, doi: 10.15537/SMJ.2023.44.8.20230230.

[102] W. A. Mumena et al., "Fast-Food Consumption, Dietary Quality, and Dietary Intake of Adolescents in Saudi Arabia," Int. J. Environ. Res. Public Health, vol. 19, no. 22, Nov. 2022, doi: 10.3390/IJERPH192215083.

[103] A. M. Albasri, M. A. Elkablawy, A. S. Hussainy, H. M. Yousif, and A. S. Alhujaily, "Impact of cyclooxygenase-2 over-expression on the prognosis of colorectal cancer patients: An experience from Western Saudi Arabia," Saudi Med. J., vol. 39, no. 8, p. 773, Aug. 2018, doi: 10.15537/SMJ.2018.8.22837.

[104] S. Alsuhaymi et al., "Untargeted metabolomics analysis of four date palm (Phoenix dactylifera L.) cultivars using MS and NMR," Nat. Prod. Bioprospect., vol. 13, no. 1, Dec. 2023, doi: 10.1007/S13659-023-00406-Y.

[105] H. A. Fadil, W. A. Sammman, and R. M. Elshafie, "Prevalence of Nonadherence to Medications among Geriatric Patients in Al-Madinah Al-Munawara City's Hospitals, Kingdom of Saudi Arabia," Int. J. Clin. Pract., vol. 2023, 2023, doi: 10.1155/2023/3312310.

[106] E. S. E. Hussein, A. M. Al-Shenqiti, and R. M. E. S. Ramadan, "Applications of Medical Digital Technologies for Noncommunicable Diseases for Follow-Up during the COVID-19 Pandemic," Int. J. Environ. Res. Public Health, vol. 19, no. 19, Oct. 2022, doi: 10.3390/IJERPH191912682.

[107] A. H. Al-Zalabani, "Adaptation and validation of the Arabic version of self-efficacy scale for mammography: A report on psychometric properties," Saudi Med. J., vol. 40, no. 7, p. 707, Jul. 2019, doi: 10.15537/SMJ.2019.7.23938.

[108] H. Kleven et al., "A neuroscientist's guide to using murine brain atlases for efficient analysis and transparent reporting," Front. Neuroinform., vol. 17, 2023, doi: 10.3389/FNINF.2023.1154080.

[109] A. F. ALAhmadi, M. F. ALSaedi, A. E. Alahmadi, M. G. Alharbi, I. H. Alharbi, and S. A. Radman Al-Dubai, "Pre-hospital delay among patients with acute myocardial infarction in Saudi Arabia. A cross-sectional study," Saudi Med. J., vol. 41, no. 8, pp. 819–827, 2020, doi: 10.15537/SMJ.2020.8.25185.

[110] A. M. Qasem Surrati, W. Mohammedsaeed, and A. B. E. Shikieri, "Cardiovascular Risk Awareness and Calculated 10-Year Risk Among Female Employees at Taibah University 2019," Front. Public Health, vol. 9, Oct. 2021, doi: 10.3389/FPUBH.2021.658243.

[111] A. Aqeeli, A. T. Alsharif, E. Kruger, M. Tennant, and H. Bakeer, "Caries prevalence and severity in association with sociodemographic characteristics of 9-to-12-year-old school children in Al-Madinah, Saudi Arabia," Saudi Dent. J., vol. 33, no. 8, pp. 897–903, Dec. 2021, doi: 10.1016/J.SDENTJ.2021.09.008.

[112] K. Perveen and N. A. Bokahri, "Comparative analysis of chemical, mineral and in-vitro antibacterial activity of different varieties of date fruits from Saudi Arabia," Saudi J. Biol. Sci., vol. 27, no. 7, pp. 1886–1891, Jul. 2020, doi: 10.1016/J.SJBS.2019.11.029.

[113] Y. A. Al-Dashti, R. R. Holt, C. L. Keen, and R. M. Hackman, "Date Palm Fruit (Phoenix dactylifera): Effects on Vascular Health and Future Research Directions," International Journal of Molecular Sciences 2021, Vol. 22, Page 4665, vol. 22, no. 9, p. 4665, Apr. 2021, doi: 10.3390/IJMS22094665.

[114] G. Gazal, H. Aljohani, K. H. Al-Samadani, and M. Z. Nassani, "Measuring the Level of Medical-Emergency-Related Knowledge among Senior Dental Students and Clinical Trainers," Int. J. Environ. Res. Public Health, vol. 18, no. 13, Jul. 2021, doi: 10.3390/IJERPH18136889.

[115] I. Ahmad, F. Alqurashi, E. Abozinadah, and R. Mehmood, "Deep Journalism and DeepJournal V1.0: A Data-Driven Deep Learning

Approach to Discover Parameters for Transportation," Sustainability (Switzerland), vol. 14, no. 9, pp. 1–70, 2022, doi: 10.3390/su14095711.

[116]R. Mehmood, "'Deep journalism' driven by AI can aid better government," The Mandarin, 2022.

[117]S. Alswedani, R. Mehmood, and I. Katib, "Sustainable Participatory Governance: Data-Driven Discovery of Parameters for Planning Online and In-Class Education in Saudi Arabia During COVID-19," Frontiers in Sustainable Cities, vol. 4, p. 871171, Jul. 2022, doi: 10.3389/frsc.2022.871171.

KnowRAG: A Zero-Shot Diagnostic Analysis of Knowledge Base Coverage in Scientific Retrieval-Augmented Generation

Assmaa MOUTAOUKKIL*, Ali EL MEZOUARY, Kaoutar BOUMALEK
IRF-SIC Laboratory-Department of Computer Science-Faculty of Science, Ibn Zohr University, Agadir, Morocco

***Abstract*—The "hallucination" problem in Large Language Models (LLMs) remains an unresolved hurdle for scientific researchers who require precise, grounded evidence. While Retrieval-Augmented Generation (RAG) aims to mitigate these errors, standard systems are often unoptimized for the structural complexities of scientific papers. We introduce KnowRAG, a zero-shot RAG pipeline specifically designed for scientific applications. Using a novel "LLM-as-a-Judge" diagnostic framework, we evaluated KnowRAG against a standalone GPT-3.5-Turbo baseline across four specialized Q&A Test Sets. Our results demonstrate that KnowRAG significantly improves factual accuracy over the baseline. More importantly, diagnostic analysis reveals that the vast majority of errors (over 46%) stem from Knowledge Base Coverage (knowledge gaps), while generation failures remain negligible at 4%. These findings suggest that retrieval and generation capabilities are no longer the primary bottlenecks in the scientific domain. Instead, this diagnostic analysis advocates for a paradigm shift from model-centric research toward expert data engineering as the definitive path to trustworthy AI. By repurposing the LLM-as-a-Judge framework as a diagnostic instrument rather than a mere performance metric, we move RAG evaluation beyond aggregate scoring toward actionable, evidence-based systemic diagnosis.**

***Keywords*—Large Language Models; Retrieval-Augmented Generation; evaluation; scientific writing; information retrieval; knowledge base; data engineering; GenAI**

I. Introduction

For a scientific researcher, the ideal assistant is a collaborator capable of synthesizing recent, multi-disciplinary literature with absolute citation fidelity. This is one of the reasons many people are getting excited about the possibilities of using Large Language Models (LLMs) [1]–[3], as LLMs have used these massive amounts of human knowledge, to completely transformed AI interaction [4]–[7] and information retrieval [2], [3], [8]–[10], and they now exhibit a very high level of capability in supply/output complex written language [2]. However, this very strength masks a profound weakness for academic research: their tendency to "hallucinate", generating seemingly convincing and persuasive texts that are subtly or blatantly inaccurate[11]–[13]. In scientific research, where the chain of evidence is vital, this error is fatal. The consequence is not just an occasional mistake, but a fundamental breach of trust, relegating these powerful tools to the status of thinking aids rather than reliable sources of knowledge. We have the linguistic engine, but it is detached from the core of the evidence and ethics. The root of this problem lies in grounding. LLM like GPT-3.5-turbo operates based on its parametric knowledge, a static, frozen snapshot of the world as perceived through its training data [14], [15]. When asked a question that goes outside the scope of this snapshot, such as the conclusions of a preprint published last month, it must generalize, often leading it to fabricate a plausible but incorrect answer. The ideal situation would be a system that seamlessly combines the inferential and linguistic capabilities of an LLM with direct, dynamic access to a reliable corpus. The dominant approach to filling this gap is retrieval-augmented generation (RAG)[5], [16]–[21]. By first retrieving relevant documents from an external knowledge base and then conditioning the LLM's generation on that specific context, RAG improves adaptability, factuality, and responsiveness to up-to-date content, and promises to link fluency with factuality[22][23][24]. The conceptual model is elegant in its separation of matters: a non-parametric memory for facts and a parametric model for reasoning and language.

However, most original work on RAG, such as the innovative framework in [19], has focused on architectural innovation, joint training of retrievers[25][26][27] and generators[28][29], or scaling to massive generic corpora. Although these are crucial advances, their evaluation has been mostly imprisoned to benchmarks such as Natural Questions, which test general faithful memorization from sources such as Wikipedia. This creates a significant and largely unexplored translation gap. Scientific literature is not a series of isolated facts; it is a dense tapestry of arguments, methodologies, results, and critiques. An improved system for retrieving a fact about a historical event may fail miserably when asked to retrieve the methodological justification for a particular experimental design, especially if the knowledge base is prepared with a generic segmentation of the text based on length, which turns a coherent argument into incoherent fragments. As a result, many existing implementations of RAG, despite their sophistication, can still produce inaccurate results in scientific contexts. The indirect consequence is a slowdown in their potential adoption; researchers cannot rely on the results without verifying them themselves exhaustively, which runs counter to the promised efficiency. This directly highlights the knowledge gap that our study attempts to address. We do not clearly understand how practical choices at the implementation level in the construction of a scientific RAG system, particularly the structuring of the knowledge base, influence its actual performance in reducing hallucinations. Previous studies have provided us with powerful engines, but little guidance on how to map their course in the specialized domain of science.

*Corresponding author

This work is based on a pragmatic conceptual model: for scientific RAG, the most effective improvements do not inevitably come from additional pre-training of massive models, but rather from an intentional and domain-appropriate design of the retrieval pipeline that feeds the generator. We build on the established RAG paradigm in accordance with the best practices mentioned in [30], but our approach differs. Instead of modifying the base LLM or training a new retrieval system from scratch, we investigate a minimal, zero-shot configuration using robust, off-the-shelf components. Our study asks a direct and practical question: Can careful knowledge base construction and retrieval setup, applied to a standard LLM, yield an impressive reduction in hallucination without any fine-tuning model?

Therefore, the objectives of this research are threefold:

- Implementation: To develop KnowRAG, a zero-shot pipeline utilizing semantic search on a processed arXiv corpus and generation via GPT-3.5-turbo.
- Evaluation: To rigorously test whether this pipeline improves factual accuracy across four domain-specific test sets compared to a standalone baseline using the LLM-as-a-Judge approach.
- Diagnosis Analysis: To perform a granular failure analysis, determine if errors stem from knowledge base gaps, retrieval misses, or generation failures.

Our primary contribution is the transformation of the LLM-as-a-Judge into a diagnostic instrument. By isolating specific systemic bottlenecks, we provide an actionable framework that moves beyond aggregate accuracy scores toward a reproducible model for reliable, inexpensive scientific assistants.

The remainder of this study mirrors our investigation process: Section II reviews related work and analyzes existing literature, while Section III details the methodology used to provide a diagnostic analysis of comprehensive RAG systems. In Section IV, we outline the experimental setup, including the creation of specialized test sets and a performance comparison against a reference LLM. Section V presents our results, offering both a quantitative evaluation of accuracy and a qualitative diagnosis of failure modes. We discuss these findings in Section VI and address the study's recommendations and limitations in Section VII. Finally, the study concludes in Section VIII.

II. Literature Review

Over the past five years, Retrieval-Augmented Generation has evolved in a field increasingly focused on algorithmic innovation [25]–[29]. Advances such as dense passage retrieval, end-to-end training of joint retriever-generator models, and more refined methods for combining retrieved content have led to measurable improvements on open-domain question answering tasks [19]. Yet this advancement in our knowledge relies that the documents used to build these models are organized, consistent, and unbiased but this is rarely true. Such a perspective cannot be upheld any longer. In narrow domains such as scientific literature, where texts contain rhetorical and informational layers and information is concentrated in different sections, the way in which the knowledge base is built becomes a defining factor for system performance. Key findings that are buried under narrative digressions, methodologies that are cut off mid-sentence, and indexing strategies that force distorted structural content into flat, indistinct lumps all highlight fundamental limitations. These are not deficiencies that can be solved by even the best-optimized retriever. Therefore, recognizing data primacy shifts the perception of the knowledge base from that of a passive archive to an active factor that shapes the maximum achievable retrieval accuracy, and consequently the factual integrity of the generated responses.

One of the greatest challenges is evaluating RAG systems as composite systems. How can we objectively assess the relevance and accuracy of a response to its sources? How can we obtain a performance measure for each component of the system? Traditional metrics from natural language processing: BLEU, ROUGE, and lexical matching do not reflect the semantic nuances of the generated text or its contextual accuracy. Human annotations are expensive and difficult to maintain throughout multiple experiments [31], [32]. The increasing use of large language models as evaluators could solve a long-standing dilemma in evaluation methodology [33]. Recent studies show that, when equipped with clear evaluation criteria and carefully designed prompts, LLMs can assess the relevance of research and the fidelity of generation, producing results that are very similar to human judgments and reaching levels of agreement comparable to those between annotators [35]. However, the significance of this change goes beyond simply replicating human evaluations effectively. The true value of the LLM as a judge lies not only in its ability to assign accurate scores but also in its ability to provide meaningful characterizations. A low relevance score in itself provides little information; what is more informative is when the model explains the failure, via chain-of-thought reasoning or dimension-specific scoring, and delivery of why relevance failed, whether due to missing content, rhetorical misalignment, or extractive fragmentation, renders visible the otherwise opaque influence of data curation decisions. While frameworks such as RAGAS [31] evaluate fidelity and accuracy, they focus on a model-centric approach: comparing retrievers in a static pipeline. This study explores a different path, using LLMs not just for final evaluation, but as tools to analyze the inner workings of the entire RAG pipeline.

This study, therefore, occupies a specific and, in our view, necessary place in this literature. Whereas previous work has optimized components in isolation or compared architectures under constant corpus conditions, we reverse the experimental logic: the algorithms are kept unchanged while the data preparation strategies are systematically modified. The goal is not to crown a single optimal configuration, but to trace, with empirical specificity, how each preservation choice propagates through retrieval and generation to shape measurable results in terms of contextual fidelity and factual accuracy. To do this, we use LLM-as-a-judge not as a complementary metric, but as a central diagnostic tool, capable of distinguishing, for example, a model that misread a passage from a model that never received enough passages to respond. This distinction, which conventional metrics cannot make, is precisely where the 'data-first' hypothesis becomes verifiable. Our implementation, KnowRAG, applies this logic in a reproducible, zero-shot pipeline designed for scientific corpora. We emphasize that this is not a proposal for a new architecture, but an argument for a

new unit of analysis. The fundamental contributions of existing RAG and their successors have equipped the field with its tools. What remains to be written, and what we aim, in part, to provide, is a systematic account of how these tools succeed or fail not in spite of the data, but because of it.

III. Methodology

This study employed a controlled experimental methodology to implement and evaluate KnowRAG, a zero-shot pipeline designed for scientific domains. Fig. 1 summarizes the KnowRAG methodology. Our approach is structured across three sequential phases to systematically measure improvements in correctness and to perform diagnostic failure analysis.

A. Knowledge Base Construction and Benchmarking

The study built a domain-specific corpus by scraping ten recent RAG-related publications from arXiv.

- Preprocessing: PDFs were converted to raw text and partitioned into 1,000-token chunks with a 30-token overlap using a recursive splitter.
- Vectorization: Each chunk was embedded using text-embedding-ada-002 and stored in a FAISS vector database for cosine similarity searching.
- Benchmarking (Q&A Test Sets): GPT-4 was utilized to generate 240 gold-standard questions (four sets of 60) directly from the corpus to ensure a verifiable "ground truth" for each query.

B. Experimental Pipeline

A comparative "Zero-Shot" procedure was executed to isolate the impact of the retrieval mechanism on a generative core (GPT-3.5-turbo). The choice of this model serves as a methodological control; by using a well-documented, stable baseline rather than a frontier model, we minimize "parametric contamination" (where the model answers from its own training memory) and more clearly isolate the performance delta attributable strictly to the RAG architecture.

- Baseline (Condition A- No retrieval): The LLM answered questions using only its internal pre-trained knowledge via a simple factual prompt.
- KnowRAG (Condition B- Retrieval-Augmented): The system embedded the query, retrieved the top-3 relevant chunks from the FAISS index, and injected them into a structured prompt. The LLM was strictly initiated to answer based only on the provided context and cite specific source IDs.
- Controls: To ensure reproducibility, parameters were fixed at a low temperature (0.1) and top_p=1.0, with no fine-tuning involved.

C. Evaluation Metric and Diagnostic Framework

To move beyond simple accuracy, we used GPT-4 as an automated judge to provide binary correctness scores and identify specific systemic bottlenecks. To mitigate evaluation circularity, we employed a hierarchical model strategy: utilizing the superior reasoning capabilities of GPT-4 to assess the outputs of the smaller GPT-3.5-Turbo generator. Following the LLM-as-a-Judge approach, GPT-4 analyzed the original question, the generated response, and the ground-truth context to determine if an answer was "correct" or "incorrect". By moving from subjective grading to objective verification, this method ensures the consistency and scale required for a deep diagnostic study. Importantly, this approach aligns closely with human expert judgment as demonstrated in [33], [34].

To understand the results, the pipeline tracked two essential metrics:

- Knowledge Base Coverage: Did the corpus contain the answer?
- Retriever Hit Rate (Top-3): Was the correct information successfully ranked within the top three documents?

These metrics are critical because they define the system's performance limits; a correct answer can only be generated if the information is both present in the corpus and successfully retrieved by the pipeline.

To understand the root cause of every error, we categorized incorrect answers into three distinct failure modes:

- Knowledge Gap: The answer was missing from the source documents entirely.
- Retrieval Failure: The information was in the corpus, but the retrieval failed to find it.
- Generation Failure: The correct context was successfully retrieved, but the LLM failed to present it faithfully.

This stepwise, instrumented methodology moves beyond reporting a simple performance delta. It provides a diagnostic framework that explains, with reasonable precision, why a zero-shot RAG system succeeds or fails in the scientific domain.

IV. Experiments

We conducted a series of controlled experiments to evaluate KnowRAG against the standalone GPT-3.5-turbo baseline. The experiments were designed to answer three core research questions. First, does retrieval augmentation significantly improve factual correctness in scientific question answering? Second, how consistent is this improvement across different types of scientific queries? And third, where do failures originate in a zero-shot RAG pipeline, and what do these failures reveal about the relative importance of knowledge base quality, retrieval accuracy, and generation fidelity?

A. Experimental Setup

All experiments were executed in a standardized computational environment. Document processing and embedding generation were performed using Python 3.10 with the LangChain, PyMuPDF, and FAISS libraries. All LLM inferences, both for generation and evaluation, were conducted via the OpenAI API. The temperature parameter was fixed at 0.1 for all generation tasks to ensure deterministic outputs, while the LLM-as-a-Judge evaluations using GPT-4 were conducted at temperature 0.0 for maximum consistency. The complete knowledge base and the generated test sets are available in the Appendix section to ensure full reproducibility.

B. Dataset and Benchmark Characteristics

Our evaluation leveraged four distinct QATestSets, each containing sixty question-answer-context triples derived from our curated arXiv corpus. The four test sets in the KnowRAG diagnostic framework were designed to vary in topic coverage to test the system's robustness across different levels of scientific complexity and domain specificity:

- QATestSet1: 4 topics, 60 questions
- QATestSet2: 7 topics, 60 questions
- QATestSet3: 6 topics, 60 questions
- QATestSet4: 4 topics, 60 questions

In each test set, the questions covered six classifications representative of scientific questions: Complex, conversational, distracting element, double, simple, and situational. For each question, GPT-4 produced a reference answer and cited the specific source identifier, thus providing verifiable truth. The final benchmark comprised 240 question-answer-context triplets, covering four distinct categories of reasoning.

To facilitate reproducibility, the four Q&A Test Sets files, including ground truth annotations, are publicly available at link.

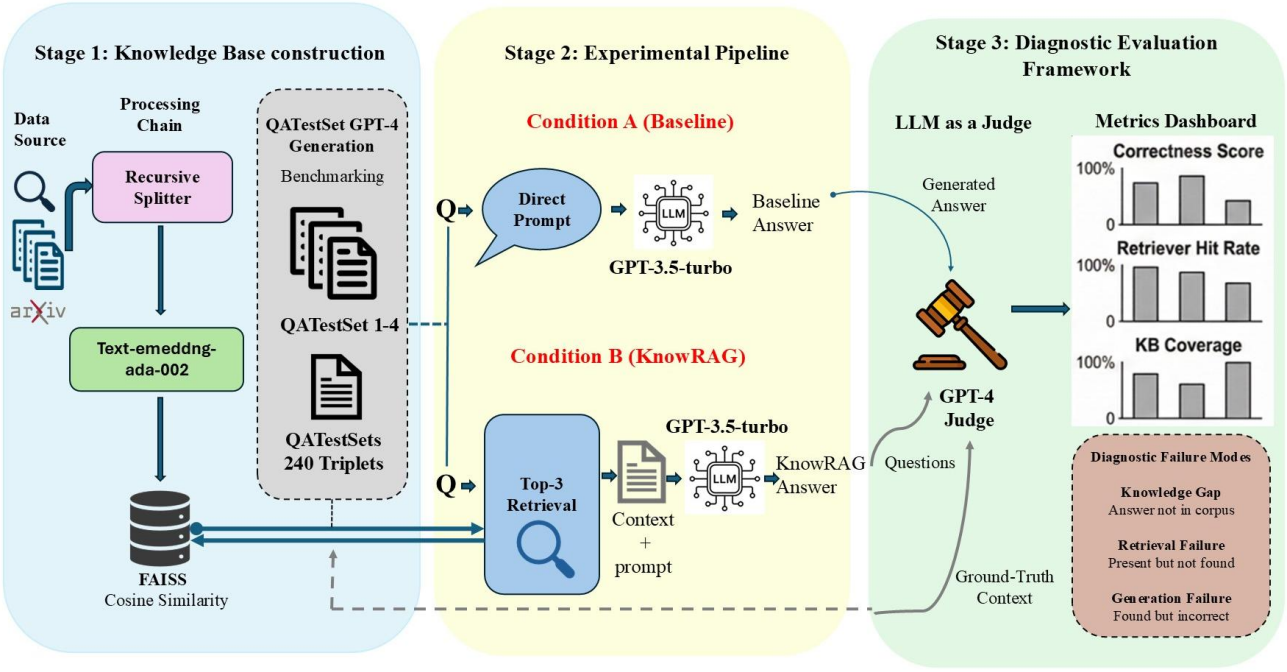


Fig. 1. The KnowRAG approach.

V. Evaluation and Results

In this section, we analyze the quantitative and qualitative performance of KnowRAG. For a detailed analysis of the evaluation Test Sets and extended diagnostic logs, please refer to Appendix A and Appendix B.

A. Comparative Performance: Baseline vs. KnowRAG

We first evaluated the factual correctness of both conditions across all four test sets. The results are presented in Table I. KnowRAG increased the average factual correctness of the baseline by more than double, from 27.5 per cent to 59.5 per cent. This improvement was consistent across all four test sets with gains ranging from 21 to 40 percentage points. The largest improvement of forty points was observed on QATestSet3, while the smallest gain of twenty-one points was observed on QATestSet2. This variation was our first indication that not all test sets, and by extension not all knowledge bases, are created equal.

TABLE I. Correctness Scores (%) by Method and Test Set Correctness is Measured via LLM-as-a-Judge (GPT-4)

Method	QATestSet1	QATestSet2	QATestSet3	QATestSet4	Average
Baseline (No Retrieval) GPT-3.5-turbo	28	29	23	**30**	27.5
KnowRAG (Retrieval-Augmented)	60	50	63	**65**	59.5
Improvement	+32	+21	+40	+35	+32

B. Diagnostic Analysis: Knowledge Base Coverage and Retriever Hit Rate

To understand why KnowRAG succeeded and where it failed, we conducted a diagnostic analysis using the metrics described in the Methodology section. Table II shows the results by Test Set.

The Knowledge Base Coverage metric reveals a catastrophic failure of our chunking strategy for certain test sets. QATestSet1

achieved excellent coverage of nearly 86%, with most of the answers to its questions preserved complete in our processed chunks. However, QATestSet2 suffered coverage of only 18%. This means that, despite the answers being present somewhere in the original PDFs, more than four out of five questions in this set had no retrievable answer after our processing.

The correlation with performance is immediate and clear. QATestSet2, which had the lowest coverage, produced the lowest correctness score (50%) and the smallest improvement over the baseline score (+21 points). In contrast, QATestSet1 and QATestSet4, with higher coverage, achieved substantially better performance. The causal relationship is clear. When the knowledge base is fragmented and incomplete, performance suffers.

With an average Retriever Hit Rate of 61%, the news report is somewhat more encouraging. When the answer was present in the knowledge base, our semantic search successfully retrieved the correct source chunk in the top three results for almost two-thirds of the questions. Performance varied, with particularly strong retrieval on QATestSet4 (70%) and weaker retrieval on QATestSet2 (45%). This suggests that methodological descriptions may be more challenging to match using cosine similarity on dense embeddings. Questions such as "What is the main contribution of paper X?" have a clear semantic center of gravity, whereas questions such as "How does the training procedure in method Y differ from standard approaches?" are more diffuse and may be harder to align with a single procedural passage.

TABLE II. INTERMEDIATE RETRIEVAL METRICS BY TEST SET (%)

Metric	QATestSet1	QATestSet2	QATestSet3	QATestSet4	AVG
Knowledge Base Coverage	85.71	18.18	44.44	66.67	53.75
Retriever Hit Rate (Top-3)	65	45	65	70	61.25

C. Failure Mode Analysis

By correlating the correctness scores with these intermediate metrics, we were able to categorize each incorrect KnowRAG response as one of three failure modes. Table III shows the resulting distribution.

The failure mode analysis yielded two profound insights that fundamentally shaped our understanding of where the real work lies in building reliable scientific RAG systems.

Firstly, the Knowledge Gap Failure is the dominant failure mode, accounting for nearly half of all errors on average and a vast 82% of errors on QATestSet2. This is not a failure of the retriever or the generator. It is a failure of the knowledge base itself. When the answer is not present in a retrievable chunk, the system has no path to correctness, regardless of the sophistication of its components. The implication is clear: the single greatest bottleneck in our zero-shot RAG pipeline is not the neural architecture of the retriever, nor the instruction-following capability of the LLM, but the mundane, understated work of deciding how to split a PDF into chunks.

Second, Generation Failure is extremely rare, accounting for only 4% of errors on average. When the correct context was successfully retrieved and provided to GPT-3.5-turbo with our structured prompt, the model almost constantly produced a faithful, accurate answer. This is a validation of the zero-shot approach. It suggests that a general-purpose instruction-tuned LLM, when given high-quality, relevant context, requires no fine-tuning to perform scientific question answering with high faithfulness. The model can read; it can follow instructions; it can synthesize. What it cannot do is conjure information that was never provided. The bottleneck is upstream.

TABLE III. FAILURE MODE DISTRIBUTION BY TEST SET (%)

Failure Mode	QATestSet1	QATestSet2	QATestSet3	QATestSet4	AVG
Knowledge Gap Failure	14.29	81.82	55.56	33.33	46.25
Retrieval Failure	20.71	36.18	19.44	16.67	23.25
Generation Failure	5.00	0.00	6.00	5.00	4.00

Retrieval Failure accounted for approximately 23% of errors, indicating meaningful room for improvement in embedding strategies, search techniques, or the choice of k. But this is, in the context of our findings, a secondary concern. Improving retrieval from 61% to 80% would be valuable, certainly, but it would still leave the system crippled if Knowledge Base Coverage remains at 53%. The order of operations matters, and our results suggest that coverage is the initial constraint.

D. Qualitative Analysis

To complement our quantitative findings, we examined specific examples from each failure category. These cases, presented in Table IV, illustrate the concrete mechanisms behind our statistical aggregates.

As illustrated in Table IV and Fig. 2 to Fig. 9, the qualitative examples confirm our quantitative results. A consistent pattern emerges: bibliographic queries regarding titles and authors fail across all test sets, despite high knowledge coverage. We attribute this to a structural bottleneck, as metadata is typically confined to specific document regions (e.g., first pages or citation blocks), making it systematically difficult to retrieve.

Furthermore, the lower Knowledge Base Coverage in Q&A Test Set 2 is directly linked to the document format. Unlike other sets, the source articles for Test Set 2 rely heavily on tabular data; because our pipeline struggled to ingest these tables, the relevant information was never indexed. This created a structural knowledge gap where the data existed in the original paper but remained absent from the system's searchable corpus.

TABLE IV. QUALITATIVE EXAMPLES OF FAILURE MODES

Question	Retrieved Context	Agent Answer	Ground Truth	Failure Mode	Key Insight
What is the title of the work by Xiaoqing Li, Jiajun Zhao, ...?	Document 105: Annual Meeting of the Association...	The title of the work by Xiaoqing Li, Jiajun Zhao... is "R2AG: Incorporating Retrieval Information into Retrieval-Augmented Generation."	The title of the work by Xiaoqing Li, Jiajun Zhao... is "Non-parametric unsupervised domain adaptation for neural machine translation."	Knowledge Gap Failure	Required paper not in corpus or metadata chunk missing.
According to Table I, which method demonstrated the highest results for the NQ dataset?	Document 22: 0.2808 0.1276 0.3882 0.2253 L...	According to Table I, the method that demonstrated the highest results for the NQ dataset is R^{2}AG (R-Squared AG).	The method R2AG+RAFT showed the highest results for the NQ dataset.	Generation Failure	Correct table retrieved but misinterpreted.
What are the three main ones? (After discussing retrieval-augmented text generation components)	Document 56: A Survey on Retrieval-Augmented T...	The three main datasets mentioned in the provided context are NQ, TQA, and HotpotQA.	The three key components of the retrieval-augmented generation paradigm are Retrieval, Augmentation, and Generation.	Retrieval Failure	Conversational thread lost; retrieved datasets instead of components.
Could you tell me the title of the academic paper by Edward J. Hu and his colleagues?	Document 119: Harrison Chase. 2022. LangChain....	I don't know.	The title of the paper by Edward J. Hu and his colleagues is "LoRA: Low-Rank Adaptation of Large Language Models."	Knowledge Gap Failure	Paper not present in corpus (low coverage).

Beyond this, three recurring failure modes stand out:

- First, knowledge gap failures often occurred not because information was absent from the corpus, but because it was fragmented by our chunking strategy. By failing to respect the original PDF structure, the strategy separated key ideas from their authors or sources.
- Second, retrieval failures occurred when embedding similarity caused semantic misalignment. For example, a question regarding training differences between REALM and DPR incorrectly retrieved a chunk on the BERT architecture. This mismatch stemmed from terminological overlap; shared terms like "encoder" and "pre-training" produced a high similarity score even though the correct passage was conceptually different. This highlights a fundamental limitation: in scientific contexts, functional relevance does not always correlate with semantic similarity.
- Third, generation failures, while rare at only 4%, stemmed from instruction-following errors. In these cases, the LLM defaulted to its internal parametric knowledge or misread the provided context, as seen in the second question of Table IV. While occasional, these instances prove that even with perfect context, accurate generation is not guaranteed. However, their low frequency confirms that modern instruction-tuned LLMs demonstrate high faithfulness when provided with relevant context.

VI. DISCUSSION

Our findings challenge the prevailing model-centric research paradigm and encourage a fundamental reorientation toward data-centric engineering in scientific RAG systems. By utilizing LLM-as-a-judge, not merely as a performance metric, but as a diagnostic instrument, we successfully isolated structural failures in retrieval coverage from cognitive failures in generation. This moves RAG evaluation toward an actionable diagnosis rather than a simple aggregate scoring.

A. Knowledge Base Coverage is the Key Obstacle

The dominant failure mode reveals a fundamental flaw in standard document processing. Standard tools, like the RecursiveCharacterTextSplitter, operate on "blind" logic, slicing text by character counts rather than rhetorical meaning. In a scientific context, this is catastrophic. A research paper is a web of connected arguments; when chunking separates a Method from its Constraint, it "break down" the knowledge. No model can reason through an answer that was physically deleted during ingestion.

B. Beyond Semantic Matching

We observe that semantic similarity is often a poor proxy for functional relevance [35]–[37]. Scientific queries often seek "differences" or "limitations". These concepts are "diffuse", meaning they are spread across paragraphs and lack the specific keywords found in the query. Dense retrievers are easily misled by "lexical noise", pulling chunks that look similar on the surface but are logically unrelated. This highlights a limitation: vectors measure how much text looks alike, not how ideas relate logically.

C. The Reliability of the LLM "Brain"

The remarkably low rate of Generation failure suggests that concerns over "hallucinations" may be misplaced in RAG systems. When provided with high-quality context, the LLM acts as a reliable synthesis engine. This implies that hallucinations are often a symptom of data absence, not a failure of logic. If the model lacks the correct data, it defaults to its own memory. Therefore, resources spent on fine-tuning models might be more effectively redirected toward expert data engineering.

D. Systematic Bias in Metadata

The consistent failure of bibliographic queries (Title/Author) reveals a positional bias. Because metadata is concentrated in specific regions like citation blocks, naive chunking splits this information into "weak" chunks that the retriever ignores. This

is not a random error; it is a systematic failure to treat document structure as a core part of the retrieval process.

These findings argue for a rebalancing of research priorities. The field has offered immense energy to improving retrievers, scaling models, and designing sophisticated fusion mechanisms. For scientific RAG, the path to reliability lies in upstream interventions. We must move away from "naive" text splitting and toward structure-aware ingestion; systems that understand a paper's layout are just as important as its texts.

VII. Limitations and Future Works

Several limitations should be considered when interpreting the findings of this study:

- Methodological circularity: Using an LLM (GPT-4) to judge a pipeline powered by an LLM (GPT-3.5) introduces potential circularity. While this study provides a diagnostic framework, these assessments have not yet been calibrated against human expert annotation.
- Generalizability: Our findings are derived from a single pipeline architecture and ten targeted articles. It remains unclear whether the "coverage-over-generation" failure mode persists across different chunking strategies, larger corpora, or specialized scientific knowledge graphs.
- To transition toward expert data engineering, future research must move beyond basic text ingestion. We suggest three concrete recommendations:
- Adopt structure-aware chunking strategies that preserve the rhetorical organization of academic writing.
- Implement hybrid retrieval approaches combining dense and sparse methods to capture functional relevance beyond semantic similarity.
- Adopt a multimodal approach that specifically addresses tabular data and equations, which were identified in Q&A Test Set 2 as a primary cause of coverage loss.

VIII. Conclusion

This study presents KnowRAG, a zero-shot retrieval-augmented generation (RAG) pipeline that achieves notably better factual accuracy compared to a Large Language Model alone. Our analysis reveals that the main limitation in performance is not retrieval or generator quality, but rather the extent of Knowledge Base coverage, with missing knowledge responsible for most errors. These results shift the focus away from scaling models, suggesting that reliable scientific AI relies more on advanced data engineering than larger models. The most impactful improvements in scientific RAG systems come from ambitious enhancements such as structure-aware document chunking, hybrid retrieval methods and more expressive knowledge representations. In essence, the effectiveness of an RAG system hinges on the quality of its knowledge base. By focusing on how we transform raw publications into retrievable knowledge, we can bridge the gap between existence in a PDF and presence in a model's response.

References

[1] D. Carmel, Y. Chang, H. Deng, and J. Y. Nie, "Future Directions of Query Understanding," Inf. Retr. Ser., vol. 46, pp. 205–224, 2020, doi: 10.1007/978-3-030-58334-7_9.

[2] C. Zhai, "Large Language Models and Future of Information Retrieval: Opportunities and Challenges," pp. 481–490, doi: 10.1145/3626772.3657848.

[3] A. Moutaoukkil and A. El Mezouary, "From Queries to Understanding: Designing the Next-Generation Search Engine," 2025 5th Int. Conf. Innov. Res. Appl. Sci. Eng. Technol. IRASET 2025, pp. 1–3, 2025, doi: 10.1109/IRASET64571.2025.11008326.

[4] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large Language Models Struggle to Learn Long-Tail Knowledge," Proc. Mach. Learn. Res., vol. 202, pp. 15696–15707, 2023.

[5] S. Wu, Y. Cui, N. Guan, and C. J. Xue, "Retrieval-Augmented Generation for Natural Language Processing : A Survey."

[6] A. Vaswani, "Attention Is All You Need," no. Nips, 2017.

[7] N. R. Mannuru, "Large Language Models (LLMs) as a Tool to Facilitate Information Seeking Behavior," 2024.

[8] Y. Zhu et al., "Large Language Models for Information Retrieval: A Survey," pp. 1–35, 2023, [Online]. Available: http://arxiv.org/abs/2308.07107.

[9] Q. Ai et al., "Information Retrieval meets Large Language Models: A strategic report from Chinese IR community," AI Open, vol. 4, pp. 80–90, 2023, doi: 10.1016/j.aiopen.2023.08.001.

[10] A. Moutaoukkil, A. Idarrou, and I. Belahyane, "Information retrieval approaches: A comparative study," Int. J. Electr. Comput. Eng. Syst., vol. 13, no. 10, pp. 961–970, 2022, doi: 10.32985/ijeces.13.10.11.

[11] Z. Xu, "Hallucination is Inevitable : An Innate Limitation of Large Language Models."

[12] Y. Zhang et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," 2023, [Online]. Available: http://arxiv.org/abs/2309.01219.

[13] Z. Li, "The Dark Side of ChatGPT: Legal and Ethical Challenges from Stochastic Parrots and Hallucination," pp. 2–4, 2023, [Online]. Available: http://arxiv.org/abs/2304.14347.

[14] T. B. Brown et al., "Language Models are Few-Shot Learners," pp. 1154–1156, 2020, [Online]. Available: http://arxiv.org/abs/2005.14165.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2018.

[16] J. Li, Y. Yuan, and Z. Zhang, "Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases," 2024, [Online]. Available: http://arxiv.org/abs/2403.10446.

[17] A. A. Khan, M. T. Hasan, K. K. Kemell, J. Rasku, and P. Abrahamsson, "Developing Retrieval Augmented Generation (RAG) based LLM Systems from PDFs: An Experience Report," pp. 1–36, 2024, [Online]. Available: http://arxiv.org/abs/2410.15944.

[18] A. Balaguer et al., "RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture," 2024, [Online]. Available: http://arxiv.org/abs/2401.08406.

[19] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Adv. Neural Inf. Process. Syst., vol. 2020-Decem, no. NeurIPS, 2020.

[20] P. Zhao, H. Zhang, Q. Yu, Z. Wang, and Y. Geng, "Retrieval-Augmented Generation for AI-Generated Content : A Survey," pp. 1–22.

[21] H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu, "A Survey on Retrieval-Augmented Text Generation," 2022, [Online]. Available: http://arxiv.org/abs/2202.01110.

[22] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks," vol. 2, pp. 296–310, 2021, doi: 10.18653/v1/2021.naacl-main.28.

[23] Y. Ke et al., "Development and Testing of Retrieval Augmented Generation in Large Language Models -- A Case Study Report," 2024, [Online]. Available: http://arxiv.org/abs/2402.01733.

[24] H. Emekci, "Maximizing RAG efficiency : A comparative analysis of RAG methods," pp. 1–25, 2024, doi: 10.1017/nlp.2024.53.

[25] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. W. Chang, "REALM: Retrieval-Augmented language model pre-training," 37th Int. Conf. Mach. Learn. ICML 2020, vol. PartF16814, pp. 3887–3896, 2020.

[26] K. Lee, M. W. Chang, and K. Toutanova, "Latent retrieval for weakly supervised open domain question answering," ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., pp. 6086–6096, 2020, doi: 10.18653/v1/p19-1612.

[27] W. Jiang, S. Zhang, B. Han, J. Wang, B. Wang, and T. Kraska, "PipeRAG: Fast Retrieval-Augmented Generation via Algorithm-System Co-design," 2024, [Online]. Available: http://arxiv.org/abs/2403.05676.

[28] H. Le et al., "FlauBERT: Unsupervised language model pre-training for French," Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc., pp. 2479–2490, 2020.

[29] L. Martin et al., "CamemBERT: a Tasty French Language Model," pp. 7203–7219, 2020, doi: 10.18653/v1/2020.acl-main.645.

[30] Y. Wu, Z. Xu, T. Shi, Z. Wang, and S. Li, "Searching for Best Practices in Retrieval-Augmented Generation."

[31] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," EACL 2024 - 18th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc. Syst. Demonstr., pp. 150–158, 2024, doi: 10.18653/v1/2024.eacl-demo.16.

[32] G. B and A. Purwar, Evaluating the Efficacy of Open-Source LLMs in Enterprise-Specific RAG Systems: A Comparative Study of Performance and Scalability, vol. 1, no. 1. Association for Computing Machinery, 2024.

[33] L. Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," no. NeurIPS, pp. 1–29, 2023, [Online]. Available: http://arxiv.org/abs/2306.05685.

[34] G. Hardy Chen, S. Chen, Z. Liu, F. Jiang, and B. Wang, "Humans or LLMs as the Judge? A Study on Judgement Bias - analiza biasów ludzkich i modeli (są inaczej zbiasowane)," pp. 8301–8327, 2024, [Online]. Available: https://github.com/.

[35] H. Zamani and W. B. Cro, "Relevance-based Word Embedding," pp. 505–514, 2017.

[36] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matching model for Ad-hoc retrieval," Int. Conf. Inf. Knowl. Manag. Proc., vol. 24-28-Octo, pp. 55–64, 2016, doi: 10.1145/2983323.2983769.

[37] K. Hui, A. Yates, K. Berberich, and G. de Melo, "PACRR: A position-aware neural IR model for relevance matching," EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc., no. July, pp. 1049–1058, 2017, doi: 10.18653/v1/d17-1110.

APPENDIX A: CORRECTNESS BY TOPICS IN FOUR QATESTSET

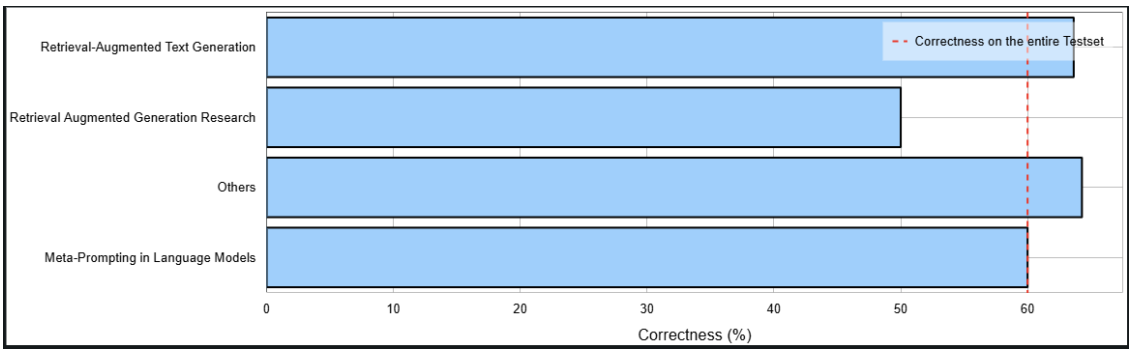


Fig. 2. Correctness by topics in the QATestSet1.

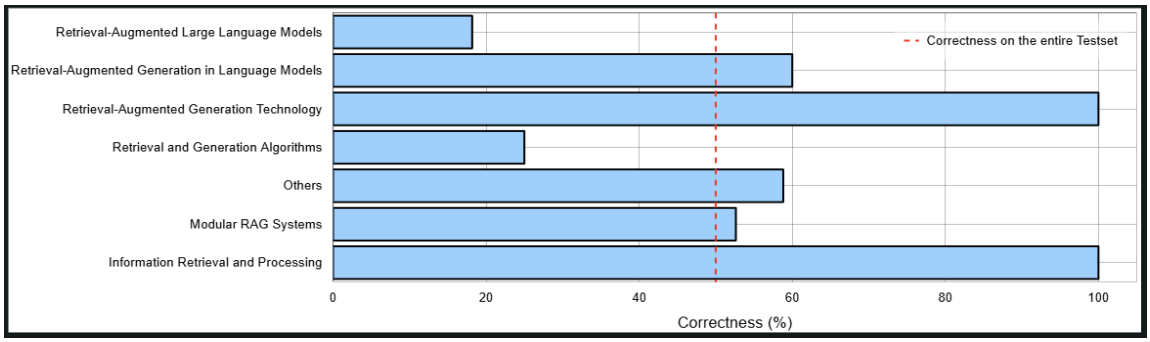


Fig. 3. Correctness by topics in the QATestSet2.

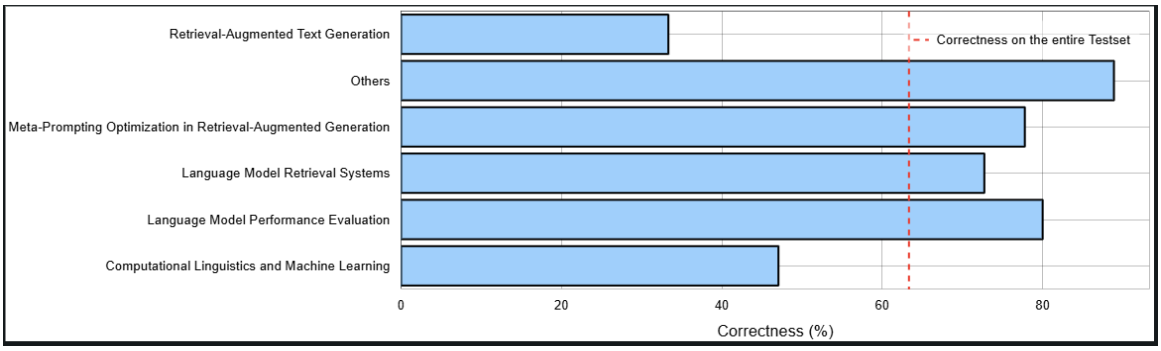


Fig. 4. Correctness by topics in the QATestSet3.

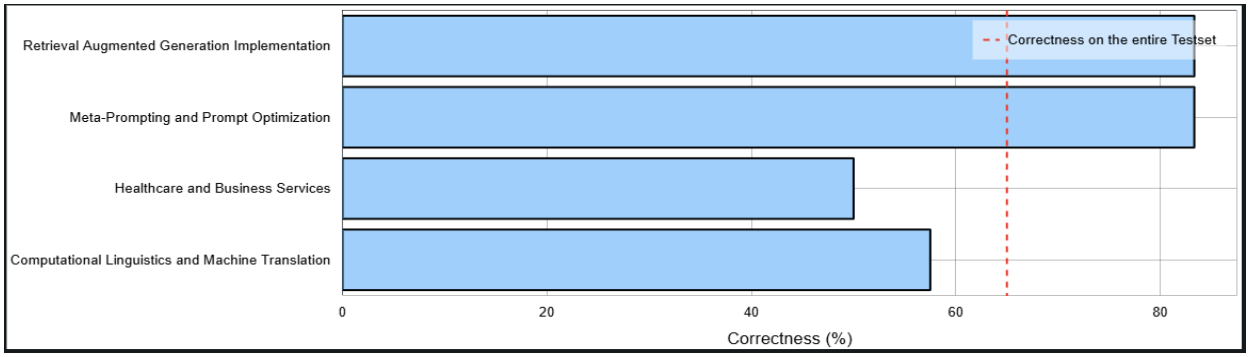


Fig. 5. Correctness by topics in the QATestSet4.

APPENDIX B: AN OVERVIEW OF THE KNOWLEDGE BASE OF FOUR QATESTSET

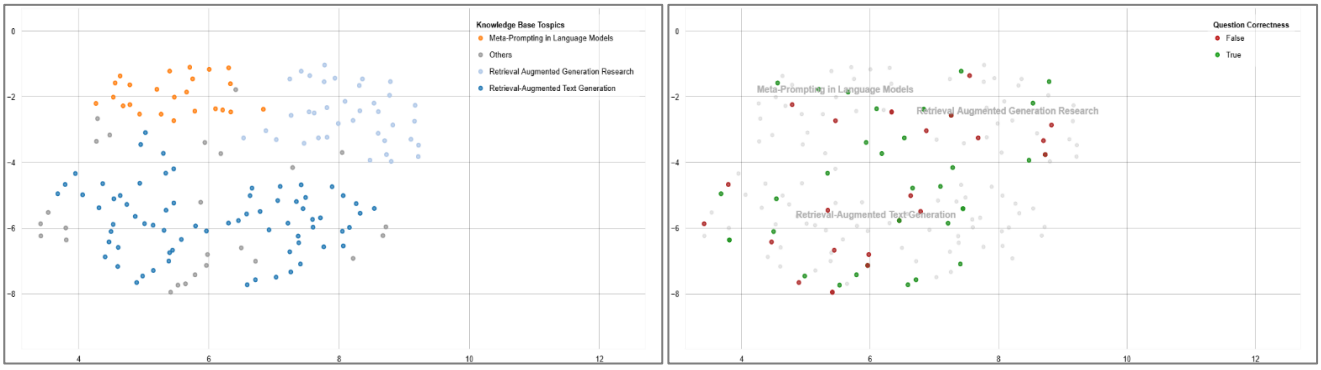


Fig. 6. An overview of the knowledge base of QATestSet1: Topics exploration and failures.

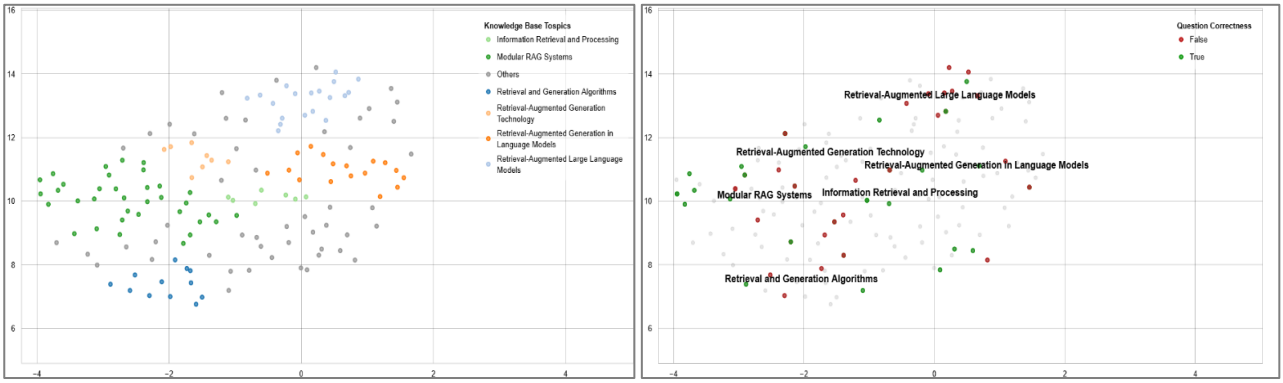


Fig. 7. An overview of the knowledge base of QATestSet2: Topics exploration and failures.

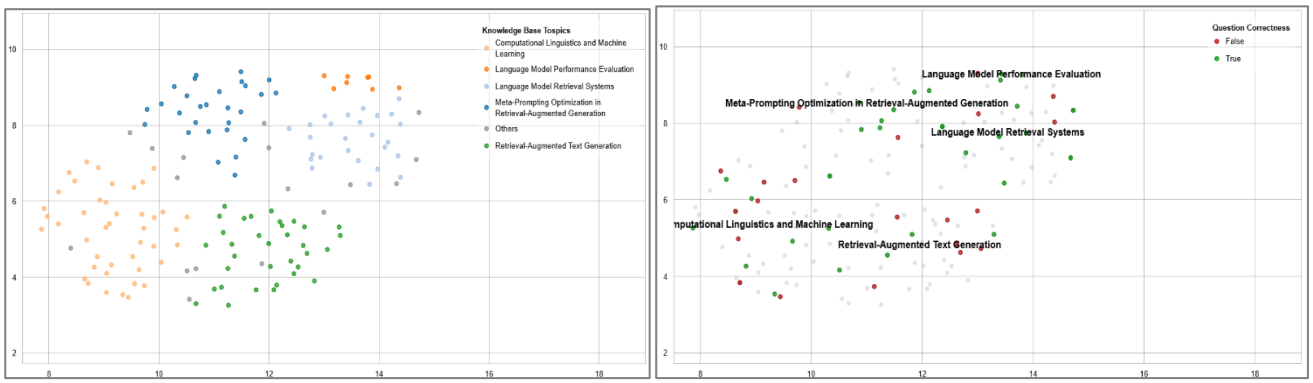


Fig. 8. An overview of the knowledge base of QATestSet3: Topics exploration and failures.

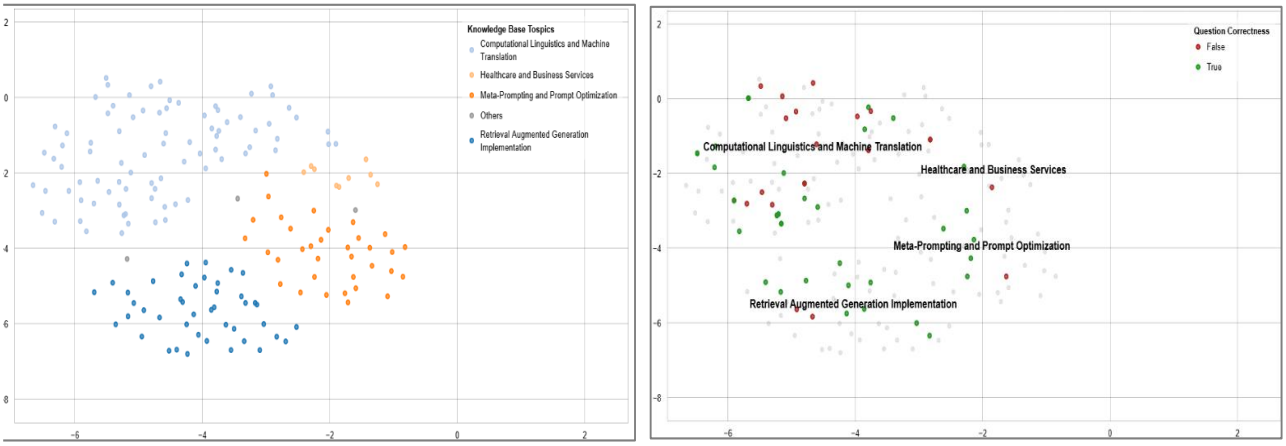


Fig. 9. An overview of the knowledge base of QATestSet4: Topics exploration and failures.

Determinants of Telemedicine Security Readiness Among ICT Professionals: Extended UTAUT Model

Fazlina Mohd Ali[1], Afraa Mohammed Mohammed Gaashan[2], Marizuana Mat Daud[3], Rimaniza Zainal Abidin[4]
Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia [1, 2, 4]
Institute of Visual Informatics, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Selangor, Malaysia [3]

***Abstract*—Ensuring strong security readiness in telemedicine is essential to protect patients, safeguard their data, and build confidence and satisfaction in today's digital healthcare platforms. Hence, this research explores and evaluates the maturity of telemedicine services from the perspectives of ICT professionals focused on security readiness. A case study with a stratified survey was conducted among ICT personnel in public hospitals in Malaysia, revealing critical gaps by examining professionals in authority for system security, data fortification, infrastructure resilience, and operational endurance of the telemedicine platform. The UTAUT model employed in this study incorporates two additional constructs, Trust in Technology and Adaptability. Trust in technology captures confidence in system safety, data privacy, and dependability. The adaptability, focused on system technical areas and user-related factors, towards secure telemedicine adoption. The SPSS Statistics Version 21 was used to analyze the data, retaining descriptive statistics, reliability testing, correlation analysis, and multiple regression. Results indicated a generally high level of security readiness among ICT professionals, with strong Trust in Technology and Adaptability to telemedicine systems. Facilitating Conditions, particularly the availability of secure infrastructure and technical safeguards, emerged as the most influential factor and significantly predicted Performance Expectancy. Although Trust in Technology and Adaptability exhibited strong positive correlations with key UTAUT constructs, their effects were statistically significant in the regression analysis, supporting all proposed hypotheses, including their influence on Satisfaction, which suggests a direct role in shaping security readiness.**

***Keywords*—Telemedicine; security readiness; ICT professionals; UTAUT; healthcare systems**

I. Introduction

Telemedicine has emerged as a vital component of digital healthcare delivery, particularly in addressing remote clinical services. However, beyond accessibility and efficiency, the security readiness of the telemedicine system has become a critical concern, especially as healthcare services increasingly rely on digital platforms for exchanging sensitive patient data. In Malaysia, the COVID-19 pandemic significantly accelerated the adoption of teleconsultation services across public and private healthcare sectors, with usage increasing sharply in primary and hospital settings [1], [2]. This was driven by the increasing need for remote consultation and outpatient care management with a reduced risk of infection.

Despite growing adoption, the success of telemedicine systems remains hindered by numerous challenges, including infrastructural limitations, security vulnerabilities, data security concerns, usability issues, and user trust in technology for secure system operations [3], [4]. These factors can hinder system performance and limit broader acceptance, mainly when not adequately addressed during implementation.

ICT (Information and Communication Technology) staff play a pivotal role in successfully deploying, maintaining, and integrating telemedicine platforms, particularly in ensuring system security, access control, data confidentiality, and resilience against operational risks. However, their perspectives are often underrepresented in policy and academic discourse, despite their experience being crucial to identifying real-world operational challenges and technical shortcomings [5].

This study employs an extended version of the Unified Theory of Acceptance and Use of Technology (UTAUT) model to assess ICT staff's perceptions of telemedicine system performance in Malaysian public hospitals. The original UTAUT framework, designed to assess technology acceptance, has been enhanced in this study by incorporating two additional constructs: trust in Technology and Adaptability, which are especially relevant in healthcare environments [6].

The primary objectives of this study are to: 1) investigate infrastructural and organizational barriers experienced by ICT personnel in supporting telemedicine services, 2) assess key predictors of system acceptance and performance from the ICT perspective, and 3) propose actionable recommendations to support more effective integration of telemedicine in Malaysia's hybrid healthcare eco-system. In this study, telemedicine security readiness refers to the organizational and technical preparedness to deploy telemedicine securely, including trusted systems, secure infrastructure, adaptable users, and effective digital workflows.

II. Related Work

Telemedicine adoption as an e-government service has been extensively studied using technology acceptance frameworks, particularly the Unified Theory of Acceptance and Use of Technology (UTAUT) [6], [7]. Venkatesh et al. [6] originally proposed UTAUT to evaluate Technology Adoption (TA), identifying Performance Expectancy (PE), Effort Expectancy (EE), and Social Influence (SI) as critical determinants. However, most UTAUT applications in healthcare focus on end users (patients, clinicians), neglecting the pivotal role of ICT staff in system implementation and in safeguarding the integrity and data protection [6], [5].

In Malaysia, the adoption of telemedicine surged during the COVID-19 pandemic, driven by policy initiatives such as the

Virtual Clinic program [5]. Studies such as [8] revealed that 66.3% of Malaysian healthcare workers viewed telemedicine as complementary to traditional care, but 29.1% resisted adoption without prior physical patient interactions, highlighting challenges related to trust and workflow integration. Notably, the perspectives of ICT workers were underrepresented, despite their critical role in addressing infrastructural barriers (e.g., cybersecurity, interoperability) [9].

Globally, UTAUT extensions have incorporated more constructs to better predict adoption in healthcare [8], [9]. For example, Marcelo et al. [10] applied UTAUT to telemedicine adoption in the Philippines, finding that adaptability and facilitating conditions (e.g., training) were key for ICT staff acceptance. Similarly, Rahman et al. [7] demonstrated that task-technology fit significantly influenced telemedicine adoption during the COVID-19 pandemic. Considering insights from existing studies provides a relevant indication that two additional constructs are required to evaluate ICT professionals' perspectives. Thus, the additional constructs added as extensions of the UTAUT model in this study are: technology trust (TT) and adaptability (AD). The relevance of the new extension of this model will be further investigated in this study to demonstrate that the new constructs are significant in evaluating the perspectives of ICT Healthcare Professionals.

There are many other related studies that contributed to the electronic medical records framework, IT capability in the Malaysian healthcare system, and the assessment of patient risks that revolve in the healthcare domain [11], [12], [13]. Despite advances contributed to in numerous studies, gaps persist in many areas of research, particularly in the Malaysian context, regarding ICT personnel in the implementation of telemedicine. Prior studies have focused on clinicians' perceptions [8] or patient satisfaction [14], with limited attention to telemedicine security assurances and the role of ICT professionals responsible for safeguarding system integrity and data protection [15]. Although trust has been examined in the context of telemedicine adoption, it is often treated as a behavioral factor rather than as a contextual factor within telemedicine's security framework. Thus, this study addresses this gap by extending UTAUT with trust and adaptability constructs [16], [5] and by exploring privacy protection and system reliability, which are central to ICT staff's responsibilities.

TABLE I. SUMMARY OF KEY RELATED STUDIES AND LIMITATIONS

Ref	Study & Focus	Limitation
[5]	Ahmad et al. (2023) ICT readiness in Malaysia	Highlights readiness gaps but lacks a theoretical model
[7]	Rahman et al. (2024) UTAUT for telemedicine (Indonesia)	Focused on general users, not backend system personnel
[8]	Khoo et al. (2023) Clinician adoption barriers in Malaysia	Only includes clinician perspectives; lacks technical staff data
[10]	Marcelo et al. (2022) ICT adoption (Philippines)	Conducted outside Malaysia; cultural/infrastructure context differs
[15]	Niu& Hong(2024) Interoperability in Malaysian telemedicine	Only examining the effect of trust in technology
[16]	Vaddhano & Tulipa (2023) Trust in UTAUT (Thailand)	Investigated trust but excluded ICT worker-specific concerns

III. METHODOLOGY

This study adopts a quantitative, cross-sectional survey design to evaluate the acceptance and perceptions of telemedicine systems among ICT professionals in selected Malaysian public hospitals. The theoretical basis for this research was the Unified Theory of Acceptance and Use of Technology (UTAUT) model, as proposed by Venkatesh et al. [1]. Two additional constructs, Trust in Technology and Adaptability, were integrated into the model based on theoretical findings, as shown in Table I. These two constructs intend to more effectively address the identified research gap and problem, particularly the limited understanding of ICT professionals' acceptance of telemedicine systems implementation in technically complex healthcare environments. These constructs capture critical aspects such as system security reliability, user confidence, and adaptability to evolving digital workflows [17], [18].

In this research, the UTAUT constructs are interpreted using telemedicine security readiness perspectives as follows:

- PE = Performance Expectancy (perceived secure performance benefit)
- EE = Effort Expectancy (secure workflow usability)
- SI = Social Influence (security culture/organizational influence)
- FC = Facilitating Conditions (security-ready infrastructure & support)
- TT = Trust in Technology (trust in system security & reliability)
- AD/A = Adaptability (adaptability to secure digital workflows)

A. Survey Instrument

The data collection tool was a structured questionnaire adapted from previously validated UTAUT-based instruments used in telemedicine and healthcare ICT. The survey was designed to assess six key constructs: Performance Expectancy (PE), Effort Expectancy (EE), Social Influence (SI), Facilitating Conditions (FC), Trust in Technology (TT), and Adaptability (AD/A). Likert-scale questions were presented to gather demographic information on participants' job roles, years of experience in ICT, and institutional affiliations, to contextualize their responses within the healthcare infrastructure.

B. Sampling and Data Collection

The case study for this research targeted ICT professionals supporting, maintaining, or deploying telemedicine systems in Malaysian public hospitals, focusing on the Klang Valley area [Hospital Kuala Lumpur, Hospital Putrajaya, Hospital Kajang, Selayang Hospital, Hospital Canselor Tuanku Muhriz UKM (HCTM)]. These hospitals were selected purposively because they were integrated into the Ministry of Health's Virtual Clinic initiative, which provided a reliable operational context for telemedicine services [4]. A purposive sampling technique was employed to ensure that only participants with relevant experience and involvement in the deployment of the telehealth system were included. The questionnaire was distributed online

using Google Forms, with access shared through hospital ICT units and administrative channels. Participation was entirely voluntary and anonymous. Respondents were informed about the study's goal, the confidentiality of their responses, and their right to withdraw at any time. This study is exploratory and context-specific, involving ICT personnel from selected public hospitals in the Klang Valley. Therefore, findings should be interpreted within this operational context.

Using Cochran's formula [19] with a 95% confidence level and a 5% margin of error, the estimated population of 90 ICT professionals from all selected hospitals yielded 69 responses (≈77% response rate), resulting in a margin of error of approximately ±5.7% at the 95% confidence level. Given the study's targeted sampling frame, high response rate, and exploratory nature, the sample was considered sufficiently representative for statistical analysis and inference [20]. For the data analysis, all obtained responses were compiled and analyzed using SPSS Statistics Version 21.

IV. Hypothesis

Research hypotheses are derived from the research purpose, which aims to examine the key determinants influencing telemedicine acceptance and its impact on the quality of care among Malaysian healthcare professionals. This study selectively examines security-relevant determinants from the extended UTAUT framework rather than testing the complete original UTAUT model. The hypotheses were tested using Pearson correlation and multiple regression analyses, which provide insight into the direction and strength of the relationships between the variables. The development of these relationships is critical for informing the implementation of healthcare settings in selected Malaysian public hospitals. Each hypothesis was evaluated using a separate regression model, based on the specified predictor-outcome relationship. The hypotheses are stated as follows:

- H1: Facilitating Conditions significantly and positively predict Performance Expectancy among healthcare professionals - Reflects the availability of secure technical infrastructure and support mechanisms [21].
- H2: Trust in Technology is positively associated with Performance Expectancy - Implies confidence in the security, reliability, and data protection of telemedicine, which increases professionals' expectations of its usefulness in healthcare delivery [22].
- H3: Adaptability of healthcare professionals is positively related to Satisfaction with telemedicine services - Indicates ICT professionals' adaptability in securing digital healthcare workflows, resulting in higher levels of satisfaction [23].

V. Results and Discussion

The analysis began with descriptive statistics to have a summary of the demographic profile of respondents and their agreement with the survey items. Pearson correlation analysis was used to examine the relationships among constructs. A multiple linear regression analysis test was performed to determine the predictors of ICT staff perceptions of telemedicine system effectiveness, including three hypothesis relationships: FC predicting PE, TT predicting PE, and AD predicting Satisfaction. All statistical procedures were conducted at a 95% confidence level, and assumptions about normality, multicollinearity, and homoscedasticity were checked and met before proceeding with the regression tests. These findings provide empirical insights into how key security-relevant factors, including system usability within secure workflows, trust in system security and reliability, infrastructure readiness, and ICT professionals' adaptability, influence telemedicine security readiness and the adoption of secure telemedicine services.

A. Descriptive Analysis

The responses were from 69, and the demographic distribution showed a balanced gender profile, with 53.6% female and 46.4% male respondents, as in Table II. The age groups were diverse, with the largest proportion being 36–45 years (30.4%), followed by 26–35 years (27.5%) and 18–25 years (20.3%). Fig. 1 shows that the highest institutional participation was from Hospital Canselor Tuanku Muhriz UKM (39.1%), highlighting its significant role in telemedicine adoption. Its partnership with the university makes it more accessible. This diverse respondent base supports the generalizability of findings related to ICT perspectives on telemedicine integration.

TABLE II. Demographic Information

Variable	Mean	Std. Deviation	Category	Frequency	Percentage
Gender	1.536	.5023	Male	32	46.4%
			Female	37	53.6%
Age	2.536	1.0513	18-25	14	20.3%
			26-35	19	27.5%
			36-45	21	30.4%
			46-55	15	21.7%

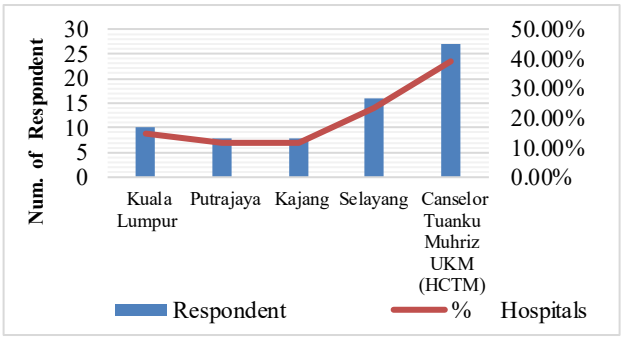


Fig. 1. Participation from ICT professionals.

Fig. 2 represents the Trust in Technology (M = 3.60) and Effort Expectancy (M = 3.60) received the highest mean scores, indicating strong confidence among ICT professionals in the security and reliability of telemedicine systems, which are essential for sustaining secure healthcare services. Social Influence (M = 3.58) and Adaptability (M = 3.58) also had positive attitudes. Performance Expectancy (M = 3.12) and Facilitating Conditions (M = 3.05) were rated comparatively

lower, indicating moderate perceptions of the system's impact on job performance and the sufficiency of infrastructure. These results offer preliminary insight into the readiness of ICT professionals to enable secure telemedicine, highlighting the importance of developing resilient technical infrastructure and performance-based benefits.

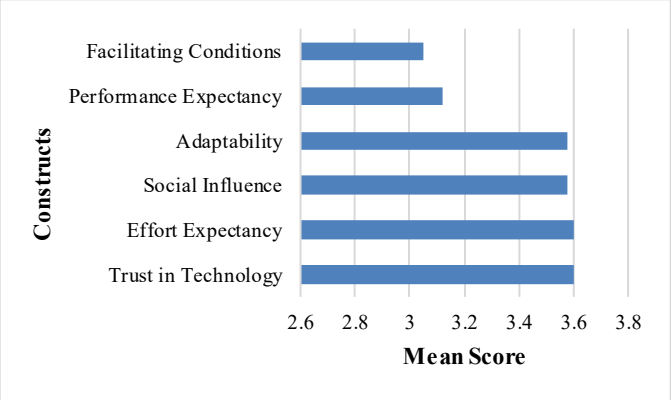


Fig. 2. Telemedicine security readiness profile (n = 69).

Overall, ICT professionals' reported readiness signals in trust and usability, while infrastructure support and perceived performance impact appeared comparatively weaker. As shown in Table III, there is room for strengthening secure telemedicine implementation.

TABLE III. DESCRIPTIVE MEAN SCORES OF TELEMEDICINE SECURITY READINESS CONSTRUCTS

Construct	Mean (M)	Security-readiness interpretation
TT	3.6	Confidence in system security & reliability
EE	3.6	Secure system perceived as usable/easy to operate
SI	3.58	Organizational support/peer influence for secure use
AD/A	3.58	Readiness to adjust to secure workflows/practices
PE	3.12	Moderate perceived impact on job performance
FC	3.05	Moderate adequacy of infrastructure/support (readiness gap)

B. Reliability Analysis

TABLE IV. RELIABILITY SUMMARY OF MEASUREMENT CONSTRUCTS

Construct	Cronbach's α	CITC range	Decision
PE	0.901	0.650–0.871	Retained (α > 0.70)
EE	0.905	0.650–0.871	
TT	0.924	0.650–0.871	
SI	0.972	0.867 – 0.909	
FC	0.972	0.782 – 0.866	
AD	0.972	0.871 – 0.895	

All constructs exceeded the recommended Cronbach's alpha threshold (α > 0.70), as shown in Table IV, indicating high reliability of the survey instrument. Corrected Item-Total Correlation (CITC) values ranged from 0.650 to 0.871, confirming each item's significant contribution to its construct. The results showed minimal variation in alpha-if-item-deleted values (0.972-0.974), confirming that no items required removal. These results validate the instrument's robustness for subsequent analysis.

C. Correlation Analysis

Pearson correlation analysis revealed significant positive relationships among all key constructs, with correlations at $p < 0.01$ (2-tailed). Notably, as in Table V, the strongest relationships were observed between EE and SI ($r = 0.894$), PE and FC ($r = 0.871$), and TT and AD ($r = 0.852$). These results are evidence that secure telemedicine readiness is closely associated with the following:

- Security-ready infrastructure and support (FC) strengthening perceived system benefit (PE)
- Trust in system security and reliability (TT) supporting ICT professionals' readiness to adapt to secure digital workflows (A).

Fig. 3 presents a heatmap of the Pearson correlations reported in Table V. Thus, the correlations suggest that the interconnected roles of usability, organizational influence, infrastructure readiness, and trust are key components in crafting secure telemedicine adoptions flexible and willing to embrace telemedicine technologies tend to report higher satisfaction.

TABLE V. PEARSON CORRELATIONS BETWEEN TELEMEDICINE SECURITY READINESS CONSTRUCTS

Cons.	PE	EE	SI	FC	TT	A
PE	1	.782**	.802**	.871**	.785**	.777**
EE	.782**	1	.894**	.708**	.817**	.779**
SI	.802**	.894**	1	.765**	.791**	.762**
FC	.871**	.708**	.765**	1	.761**	.752**
TT	.785**	.817**	.791**	.761**	1	.852**
A	.777**	.779**	.762**	.752**	.852**	1
**. Correlation is significant at the 0.01 level (2-tailed).						

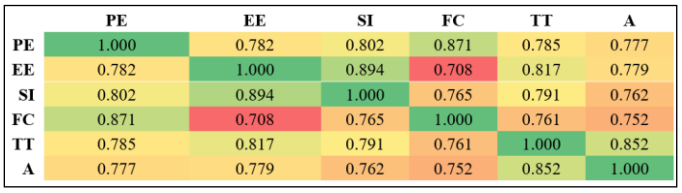


Fig. 3. Pearson correlation heatmap of telemedicine security readiness constructs.

D. Regression Analysis

Multiple regression analysis further supported these findings through three separate predictive models, aligned with the study hypothesis. The regression outcomes are summarized in Table VI. Model 1 revealed that Facilitating Conditions significantly predicted Performance Expectancy ($\beta = 0.483$, $p = 0.007$), underscoring the importance of a security-ready infrastructure, reliable system support, and technical safeguards in shaping ICT professionals' perceptions. Model 2 confirmed that Trust in Technology also positively influenced Performance Expectancy ($\beta = 0.428$, $p = 0.012$). Meanwhile, Model 3

identified Adaptability as a significant predictor of user Satisfaction ($\beta = 0.521, p < 0.001$), highlighting its importance in the successful adoption of secured telemedicine systems. Collectively, these findings highlight that a supportive technological environment and adaptable ICT staff are crucial drivers of secure telemedicine adoption.

TABLE VI. MULTIPLE REGRESSION RESULTS FOR TELEMEDICINE SECURITY READINESS.

Model	Dependent Variable	Significant Predictor(s)	Standardized β	*p*-Value
Model 1	Performance Expectancy (PE)	Facilitating Conditions (FC)	0.483	0.007
Model 2	Performance Expectancy (PE)	Trust in Technology (TT)	0.428	0.012
Model 3	Satisfaction	Adaptability (AD/A)	0.521	< 0.001

E. Hypothesis Test

The hypothesis testing results revealed that Facilitating Conditions significantly predicted Performance Expectancy among ICT professionals, strongly supporting H1. In contrast, both Trust in Technology (H2) and Adaptability (H3) significantly predicted Performance Expectancy, while Adaptability showed a strong positive correlation with Performance Expectancy. However, it did not emerge as a significant predictor of Performance Expectancy in the regression models, yet it was significant in predicting satisfaction. These findings suggest that while trust and adaptability are valued, infrastructure and technical support remain the most influential factors in shaping ICT professionals' perceptions of telemedicine effectiveness. Thus, H1 is fully supported, while H2 is supported, and H3 is supported for Satisfaction but not for Performance Expectancy. The hypothesis testing summary is presented in Table VII.

TABLE VII. HYPOTHESES ANALYSIS

Hypothesis	Relationship Tested	Correlation (r)	Regression (β)	*p*-value	Supported?	Addresses Objective
H1	FC → PE	0.871**	0.483	0.007	Yes	Security-ready infrastructure/support boosts perceived usefulness
H2	TT → PE	0.785**	0.428	0.012	Yes	Trust in system security & reliability boosts perceived usefulness
H3	AD → Satisfaction	0.786**	0.521	<0.001	Yes	Adaptability to secure workflows increases satisfaction
*p < 0.01 (2-tailed). (** indicates significant correlation at 0.01 level.)*						

TABLE VIII. TELEMEDICINE ANALYSIS SUMMARY

Evidence (from analysis)	Key finding	Security-readiness implication
Descriptive Statistics (PE, EE, SI, FC, TT, AD)	Mean range: 3.05–3.60; highest means: TT = 3.60, EE = 3.60; lower means: PE = 3.12, FC = 3.05	ICT professionals demonstrate a positive readiness toward secure telemedicine; however, security-ready infrastructure and support, as well as perceived performance benefits, require further strengthening.
Reliability Analysis (23 items)	Overall reliability: Cronbach's Alpha = 0.974	The measurement instrument is robust, supporting confidence in the reported telemedicine security readiness findings.
Correlation Analysis	TT–PE: r = 0.850, AD–Satisfaction: r = 0.786 ($p < 0.01$)	Trust in system security/reliability is strongly linked to perceived usefulness, while adaptability is strongly linked to satisfaction—both are key indicators of readiness.
Regression Model 1	FC → PE: β = 0.483, p = 0.007	Security-ready infrastructure, support, and safeguards significantly strengthen ICT professionals' perception of telemedicine usefulness.
Regression Model 2	TT → PE: β = 0.428, p = 0.012	Higher trust in system security and reliability increases perceived usefulness, supporting the readiness of secure telemedicine.
Regression Model 3	AD → Satisfaction: β = 0.521, p < 0.001	ICT professionals' adaptability to secure digital workflows significantly improves satisfaction and supports sustained secure adoption.

Summary in Table VIII outlines the key findings from the analysis, showing strong support for H1 and partial support for H2 and H3. These results suggest that telemedicine success depends not only on usability but also on organizational preparedness to support secure digital healthcare services, aligning with broader telemedicine security governance objectives.

VI. CONCLUSION AND FUTURE WORK

The study provides empirical evidence on the security readiness of telemedicine from the perspectives of ICT professionals supporting telemedicine services in selected public hospitals, particularly in the Malaysian case study. The overall perception was positive, especially for Trust in Technology and Effort Expectancy, indicating that secure telemedicine platforms are generally accepted as reliable and adaptable. However, a similarly low perception of Performance Expenditure and Facilitation Conditions suggests that the perceived benefits of telemedicine and the adequacy of the infrastructure remain areas for improvement. A summary of key results is provided in Table VIII.

The inferential findings further prove the key readiness drivers for secure telemedicine. Whereby, correlation values indicate a strong interrelationship among the constructs, while regression results confirm that security-ready infrastructure and support (Facilitating Conditions) and trust in system security and reliability (Trust in Technology) both contribute to a gradual improvement in perceived usefulness (Performance Expectancy). Additionally, Adaptability also significantly predicts Satisfaction, highlighting the importance of ICT professionals' readiness to operate within a secure digital work process. Collectively, these findings suggest that secure telemedicine adoption depends not only on system usability but

also on organizational preparedness, trusted operations, and workforce adaptability, offering practical direction for robust secure digital healthcare services.

Extended studies in this research area can be conducted by employing a larger, more diverse sample of ICT professionals from various regions and hospital levels across Malaysia, including private and rural health facilities. As this study used correlation and multiple regression for exploratory purposes and with a limited sample size, future studies with larger samples may employ SEM to examine mediation and full model fit. This will make the results more generalizable and allow for comparisons across various healthcare settings. Longitudinal studies are also recommended to quantify how perceptions and uptake patterns evolve.

Moreover, future studies may extend this work by examining telemedicine cybersecurity practices, security governance frameworks, and regulatory compliance models to further strengthen telemedicine security readiness across healthcare systems. Interpreting the impact of cybersecurity, digital literacy, and departmental collaboration on telemedicine deployment would help facilitate a more comprehensive understanding of its success.

ACKNOWLEDGMENT

The Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (WARISAN-2025-004) funds this research. Heartfelt appreciation is expressed for the funding and facilities provided, which were essential for completing this study.

REFERENCES

[1] Lim, N.A., Mohamad, N., Abdullah, M.: Exploring the Rise of Telehealth Services in Malaysia: A Retrospective Study. Journal of Science, Technology and Innovation Policy 9(2), 1–12 (2023).

[2] Ng, S.W., Hwong, W.Y., Husin, M., Ab Rahman, N., Nasir, N.H., Juval, K., Sivasampu, S.: Assessing the availability of teleconsultation and the extent of its use in Malaysian public primary care clinics: a cross-sectional study. JMIR Formative Research 6(5), e34485 (2022).

[3] Wahab, S.F.A., Ismail, N.A., Wahid, N.A.A., Ali, M.A.R.M., Ismail, A.R.: Challenges in telemedicine implementation in Malaysian public hospitals: a qualitative study. BMJ Open 13(3), e067210 (2023).

[4] Hussein, A., Goh, H.T., Kamaruddin, A.: Factors influencing telehealth adoption among consumers in Malaysia: trust, usability, and UTAUT extension. Systems 11(2), 92 (2023).

[5] Tan, S.-H., Wong, C.-K., Yap, Y.-Y., Tan, S.-K.: Factors influencing telemedicine adoption among physicians in the Malaysian healthcare system: a revisit. BMJ Open 14(1), e078508 (2024).

[6] Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly 27(3), 425–478 (2003).

[7] Rahman, B., Prawitasari, A.P., Setiawan, Y.A., Wijaya, L.: Telemedicine Application Adoption During the COVID-19 Pandemic: The Lens of the UTAUT Framework Model. Journal of Applied Research in Computer Science and Information Systems 2(1), 121–131 (2024).

[8] Khoo, A., Hassan, M., Karim, R., Lim, S.: Evaluating Telemedicine Perception and Readiness Among Healthcare Workers in Malaysia. Perspectives in Health Information Management 20(3), 1f (2023).

[9] Shrestha, N.M., Alsadoon, A., Prasad, P.W.C., Hourany, L., Elchouemi, A.: Enhanced e-health framework for security and privacy in healthcare system. In: 2016 International Conference on Digital Information Processing and Communications (ICDIPC), pp. 75–79. IEEE (2016).

[10] Marcelo, P., Ylagan, J., Razon, N., Arce, P.: Factors Affecting Adoption of a Telemedicine Device by Primary Care Health Workers in the Philippines. Acta Medica Philippina 56(11) (2022).

[11] A. Alhammad, M. M. Yusof, and D. I. Jambari, "Assessing Cloud Computing Adoption in the Malaysian Healthcare Industry based on IT Capability," Expert Rev. Med. Devices, vol. 21, no. 3, pp. 217–229, 2024, doi: 10.1080/17434440.2024.2315024.

[12] N. M. Noh, N. S. Mohd Satar, H. Sallehudin, N. A. Abu Bakar, A. F. Md Fadzil, and F. Yahya, "Assessing Cloud Computing Adoption in the Malaysian Healthcare Industry based on IT Capability," in 2025 International Conference on Electrical Engineering and Informatics (ICEEI), 2025, pp. 1–6. doi: 10.1109/ICEEI68459.2025.11331049.

[13] N. Rohani and M. M. Yusof, "Unintended consequences of pharmacy information systems: A case study," Int. J. Med. Inform., vol. 170, p. 104958, 2023, doi: https://doi.org/10.1016/j.ijmedinf.2022.104958.

[14] Othman, E., Al-Amry, N., Wong, A.: Patient satisfaction with teleconsultation during COVID-19 pandemic: a descriptive study for mental health care in Malaysia. Malaysian Journal of Public Health Medicine 22(1), 100–109 (2022).

[15] Niu, L., Hong, J., Ma, L.: Examining the effect of trust in technology on healthcare professionals' telemedicine use intention: a structural model approach. Journal of Biomedical Informatics 146, 104426 (2024).

[16] Vaddhano, N., Tulipa, D.: Impact of IT identity and data privacy on mobile telemedicine use: A UTAUT perspective. Journal of Mobile and Biomedical Informatics 5(2), Article 9 (2023).

[17] M. M. Yusof, T. Takeda, Y. Shimai, N. Mihara, and Y. Matsumura, "Evaluating health information systems-related errors using the human, organization, process, technology-fit (HOPT-fit) framework," *Health Informatics J.*, vol. 30, no. 2, 2024, doi: 10.1177/14604582241252763.

[18] M. M. Yusof, Y. Shimai, and T. Takeda, "Safe Medication: An Integrated Approach," Stud. Health Technol. Inform., vol. 327, pp. 422–423, 2025, doi: 10.3233/SHTI250367.

[19] W. G. Cochran, *Sampling Techniques*, 3rd ed. New York, NY, USA: Wiley, 1977.

[20] Fink, A.: Conducting Research Literature Reviews from the Internet to Paper. Thousand Oaks, CA: Sage Publications (2019).

[21] J. C. Nunnally and I. H. Bernstein, Psychometric Theory, 3rd ed. New York, NY, USA: McGraw-Hill, 1994.

[22] Rajashekhar, S.A., Sharma, K., Trivedi, M.K., Villagómez Galindo, M., Singh, K., Valencia, A.B.M.: Investigating the impact of adoption of smart healthcare technologies on quality of patient care using a two-staged SEM-ANN approach. Discover Public Health 22, article 350 (2025).

[23] Kung, L.-H., Yan, Y.-H., Kung, C.-M.: Exploring Telemedicine Usage Intention Using Technology Acceptance Model and Social Capital Theory. Healthcare (Basel) 12(13), 1267 (2024).

Evaluating Perceptual Reliability of Latent Attribute Control in Diffusion-Based Fashion Generation

Noriaki Kuwahara[1], Shintaro Kawanami[2], Takashi Sato[3], Dongeun Choi[4]*
Kyoto Institute of Technology, Kyoto, Japan[1, 2, 4]
Tokyo Denki University, Tokyo, Japan[3]

***Abstract*—Although diffusion-based image generation models enable high-quality synthesis of fashion images, the reliable control of perceptual attributes in these models remains poorly understood. Current evaluation approaches primarily rely on semantic similarity metrics, such as CLIP scores, which may not accurately reflect human perceptual judgments. This study proposes a three-layer evaluation framework linking latent space geometry, semantic embedding space, and human perception. First, latent attribute directions are validated using geometric quality-control metrics measuring linearity and centrality. Second, semantic consistency is examined through directional projection in CLIP embedding space. Third, a two-alternative forced-choice experiment is conducted with 37 participants, and perceptual strength is estimated using a Bradley-Terry preference model. Experiments cover gender and garment conditions for four fashion attributes: fit, lightness, glossiness, and pattern scale. Results reveal that fit exhibits strong cross-layer alignment, while pattern scale shows semantic and perceptual ambiguity. The findings highlight that perceptual reliability in controllable generation is attribute-dependent and that semantic metrics alone cannot fully replace human evaluation.**

***Keywords*—Diffusion models; controllable generation; latent space analysis; human preference modeling; perceptual reliability; fashion image generation**

I. Introduction

Diffusion-based image generation models demonstrated remarkable visual quality and are widely used in text-to-image synthesis [1], [2]. In fashion applications, these models enable controllable manipulation of garment appearance through latent-space operations. However, reliable control of perceptual attributes such as fit, lightness, glossiness, and pattern scale has not been sufficiently validated.

Current evaluation methods primarily rely on semantic similarity metrics, including CLIP-based scores or reward models trained on human preferences [3], [4]. These approaches measure alignment between generated images and textual prompts, but do not explicitly assess whether latent attribute manipulations are geometrically stable or perceptually consistent. As a result, latent representation behavior, semantic embedding responses, and human perceptual discernment are often treated interchangeably.

Previous studies show that semantic attributes may correspond to specific directional changes in latent space [5], [6]. Despite these findings, the geometric validity of these directions and their perceptual reliability within diffusion models have not been systematically analyzed. Furthermore, while human preference modeling is increasingly employed in generative AI evaluation [4], [7], its relationship to latent space geometry and semantic embedding behavior is still unclear.

To address this gap, we propose a three-layer evaluation framework that links latent space geometry, semantic embedding space, and human perception. Our method employs a latent direction quality-control mechanism to validate the linearity and centrality of attribute directions prior to semantic and human evaluation. Experiments under gender and garment conditions measured performance across four perceptual attributes. Cross-layer consistency analysis shows that the perceptual reliability of controllable diffusion generation is attribute-dependent and cannot be fully predicted by semantic similarity metrics alone.

The contributions of this study are:

- A three-layer evaluation framework separating latent geometry, semantic consistency, and human perceptual reliability.
- A latent direction quality control approach for validating geometric stability.
- An empirical analysis revealing attribute-dependent alignment across evaluation layers.

II. Related Work

A. Controllable Diffusion Models

Diffusion-based generative models are widely known for producing high-fidelity images [1], [2]. Their iterative denoising framework supports stable training dynamics and allows for flexible conditioning. While many diffusion models focus on text-conditioned generation, some have introduced additional control signals into their architectures to guide structural or semantic properties. For example, ControlNet introduced extended diffusion models with conditional control branches to improve structural consistency under user constraints [8]. More recent work has further explored efficient and scalable virtual try-on systems using knowledge distillation techniques, improving both realism and computational efficiency [9]. In fashion-related applications, diffusion-based virtual try-on systems have demonstrated realistic garment transfer while preserving the human form. Most of these systems focus on achieving realistic results and maintaining structural coherence. However, limited attention has been paid to systematically evaluating the controllability of perceptual attributes.

*Corresponding author

Although diffusion models allow manipulation of latent spaces, the geometric properties of attribute directions in diffusion latent spaces have not been extensively examined. Unlike deterministic GAN-based architectures, diffusion models involve stochastic denoising processes. This presents challenges in maintaining stable latent attribute directions. As a result, controllability depends on explicit validation of the directions of geometric attributes rather than on assuming their stability.

B. Latent Space Geometry and Directional Editing

Latent space interpretability has been widely studied in GAN frameworks. StyleGAN demonstrated that semantically meaningful transformations can emerge along structured latent representations [10]. However, prior work has also challenged common assumptions in disentangled representation learning, indicating that such latent structures may not be as robust or universally interpretable as often assumed [11]. InterFaceGAN further showed that certain attributes correspond to approximately linear directions in latent space, enabling semantic editing through vector operations [5]. Subsequent studies emphasized the importance of geometric alignment and stability in defining controllable directions [6], [12].

Representation learning theory suggests that meaningful variation predominantly lies in structured subspaces rather than arbitrary directions [13]. While most analyses focus on GAN architectures that employ deterministic latent-variable mappings, diffusion models differ significantly due to their latent characteristics arising from stochastic noise injection and iterative refinement. Consequently, the geometric validity, centrality, and stability of attribute directions in diffusion latent space are underexplored and require dedicated evaluation mechanisms.

C. Semantic Embedding-Based Evaluation

Semantic alignment between generated images and corresponding textual descriptions is commonly evaluated using CLIP-based similarity metrics [3]. Embedding-based reward models, such as ImageReward, are designed to approximate human preference through learned scoring mechanisms [4]. These evaluation approaches provide scalable, automatic assessment and are now standard benchmarks for text-to-image generation.

However, distance-based similarity metrics do not explicitly separate intended attribute variation from unrelated visual changes. Existing literature on evaluation has shown that global error or similarity measures may not capture meaningful structural features of human perception [14]. Recent studies on human preference modeling in generative systems further emphasize that embedding-based metrics only partially correlate with human discretion [15]. For this reason, direction-sensitive analysis in embedding space is needed to verify that attribute progression is consistently preserved, not just semantic similarity increased.

D. Human Preference and Perceptual Reliability

Human evaluation remains fundamental to assessing the inherently subjective aspects of generative models. Human preference modeling systems typically employ experimental methods, such as two-alternative forced-choice (2AFC) and pairwise comparison frameworks, to statistically quantify human subjective discernment [7]. The resulting comparative data is then analyzed using the Bradley-Terry model [16] to construct latent preference scales that reflect human preferences. These modeling processes not only play a central role in evaluation but also serve as a key component for aligning reward systems within modern generative AI pipelines [4], [15].

Despite the growing adoption of human preference modeling techniques, most existing studies evaluate either automatic semantic alignment or human preference independently. They have not yet examined the relationship among latent geometry, semantic embedding behavior, and perceptual reliability collectively. A consolidated framework integrating geometric validation, semantic direction analysis, and human perceptual assessment is necessary to facilitate structured interpretation and understanding of controlled generation aligned with human factors.

Recent studies have increasingly emphasized the importance of controllability and systematic evaluation in generative AI systems [18], [19], highlighting the limitations of relying solely on semantic similarity metrics. These trends further motivate the need for a unified framework that systematically links latent space manipulation, semantic embedding behavior, and human perception.

III. Proposed Framework

A. Problem Formulation

Recent diffusion-based image generation models provide continuous manipulation of visual attributes in the latent space. However, controllable latent directions do not guarantee perceptual reliability.

In this study, we distinguish between three conceptually distinct layers:

- Latent geometry: the internal representation of the generative model.
- Semantic embedding space: a machine-interpretable meaning structure (e.g., CLIP).
- Human perception: subjective discernment of attributes.

Controllable generation implicitly assumes that a given perceptual attribute varies and can be approximated along a one-dimensional axis in the latent space. However, this geometric construction implicitly assumes that:

- Attribute variation is locally linear in latent space.
- The base image lies near the midpoint of the attribute axis.
- Progression along the axis corresponds to monotonic perceptual change.

These assumptions carry significant and non-trivial implications.

From a geometric standpoint, high-dimensional generative latent spaces do not ensure perfectly aligned semantic subspaces.

From a psychophysical standpoint, perceived attribute intensity is generally non-linear with respect to changes in

physical stimulus, as suggested by classical laws such as Weber–Fechner.

Perceptual reliability is, therefore, defined as cross-layer monotonicity and not just latent controllability. This approach ensures the ordering induced by the generation parameter α is preserved in both semantic and human perceptual evaluations.

More formally, let:

- α denote the generation parameter along the latent direction,
- $s(\alpha)$ denotes the semantic projection score,
- $\theta(\alpha)$ denotes the perceptual strength estimated from human evaluation.

An attribute manipulation is considered perceptually reliable if:

- The latent configuration satisfies geometric consistency,
- $s(\alpha)$ is monotonic with respect to α,
- $\theta(\alpha)$ is monotonic with respect to α,
- The ordering induced by $s(\alpha)$ and $\theta(\alpha)$ is positively correlated.

Thus, the central research question is: What geometric conditions enable latent attribute direction to preserve monotonicity through semantic embedding and human perception?

This formulation transforms attribute control from a generation problem into a cross-layer consistency problem.

B. Latent Attribute Direction Hypothesis

Let $z_0 \in \mathbb{R}^d$ denote the latent representation of a base image generated by the diffusion model.

For a given perceptual attribute a, we obtain two endpoint latent codes which are intended to represent the minimal and maximal perceptual strength of attribute a.

$$z_{a,\min}, z_{a,\max}$$

We define the latent attribute direction as Eq. (1):

$$\mathbf{d}_a = z_{a,\max} - z_{a,\min}. \quad (1)$$

To ensure symmetry and interpretability, we adopt a midpoint-centered and normalized parameterization.

The midpoint is defined as Eq. (2):

$$z_{a,mid} = \frac{z_{a,\max} + z_{a,\min}}{2} \quad (2)$$

The normalized direction is Eq. (3):

$$\hat{\mathbf{d}}_a = \frac{\mathbf{d}_a}{\| \mathbf{d}_a \|}. \quad (3)$$

The latent attribute axis is then expressed as Eq. (4):

$$z(\alpha) = z_{a,mid} + \alpha \frac{\| \mathbf{d}_a \|}{2} \hat{\mathbf{d}}_a, \qquad \alpha \in \mathbb{R} \quad (4)$$

whereby,

$$\alpha = \pm 1$$

correspond to the endpoints of the minimal and maximal attributes.

This formulation separates two properties:

- Axis validity — whether $z_{a,\min}$ and $z_{a,\max}$ define a coherent one-dimensional direction.
- Base centrality — whether z_0 lies near $z_{a,\text{mid}}$.

We therefore formulate the following hypothesis:

Hypothesis H_1:

The triplet $(z_{a,min}, z_0, z_{a,max})$ is approximately collinear in the ambient latent space $\mathbb{R}^d$, and displacement along the normalized direction $\hat{\boldsymbol{d}}_a$ preserves perceptual ordering.

If H_1 holds, varying α should induce a systematic perceptual change in attribute a without introducing unintended semantic drift. Because diffusion latent spaces are high-dimensional and stochastic, H_1 is not guaranteed a priori and must be empirically validated.

In the next section, we introduce geometric criteria for testing H_1.

C. Geometric Quality Control in Latent Space

1) Linearity deviation (Δ): To evaluate hypothesis H_1, we first introduce a geometric measure that quantifies deviation from ideal midpoint symmetry.

The midpoint of the attribute axis is defined by Eq. (2). Assuming ideal symmetry and perfect collinearity, the base latent representation coincides with the midpoint.

Equivalently,

$$z_{a,\max} + z_{a,\min} - 2z_0 = 0.$$

We, therefore, define the linearity deviation as Eq. (5):

$$\Delta = \| z_{a,\max} + z_{a,\min} - 2z_0 \| \quad (5)$$

Geometric interpretation:

- $\Delta = 0$ indicates perfect midpoint symmetry.
- Larger Δ indicates deviation from collinearity.
- Non-zero Δimplies geometric distortion of the attribute axis relative to the base configuration.

Notably, Δ evaluates structural consistency exclusively in latent space, independent of semantic or perceptual interpretation.

2) Centrality (c): While Δ measures absolute deviation, it depends on the scale of the attribute span. To obtain a scale-invariant quantity, we define centrality as Eq. (6):

$$c = 1 - \frac{\| z_{a,\max} + z_{a,\min} - 2z_0 \|}{\| z_{a,\max} - z_{a,\min} \|} \quad (6)$$

Properties:

$c \approx 1 \rightarrow$ base is near the midpoint.

$c \approx 0 \rightarrow$ base is near one endpoint.

$c < 0 \rightarrow$ base lies outside the attribute span.

Thus, centrality quantifies the extent to which the base image is centered relative to the attribute axis.

From a geometric perspective, c quantifies the symmetry of displacement. In psychophysical terms, poor centrality may induce asymmetric perceptual sensitivity along the attribute axis.

3) QC decision principle: Geometric validation must take place before semantic or perceptual evaluation.

We define a quality control (QC) mechanism based on Δ and c. An attribute direction is classified as:

- PASS if Δ is sufficiently small and c is sufficiently high.
- WEAK-PASS if only one condition is moderately satisfied.
- FAIL if both conditions indicate significant geometric distortion.

The specific numerical thresholds are determined empirically and reported in the experimental section.

Conceptually, QC establishes a necessary geometric condition for ensuring perceptual reliability:

- If geometric symmetry fails, cross-layer monotonicity cannot be interpreted meaningfully.
- If QC is satisfied, subsequent semantic and perceptual evaluations assess whether geometric consistency propagates across layers.

Thus, geometric QC constitutes the structural foundation of the proposed three-layer evaluation framework.

D. Semantic Direction Validation (clip)

Section III-C establishes geometric consistency in latent space. To ensure perceptual reliability, the attribute progression should be preserved in the semantic embedding space.

We therefore analyze whether the latent attribute axis defined in Section III-B remains consistently oriented in the CLIP embedding space.

Let Eq. (7):

$$\mathbf{v}_i = f_{\text{img}}(x_i) \in \mathbb{R}^D \tag{7}$$

denote the CLIP image embedding of the generated image x_i, where D denotes the dimensionality of the CLIP embedding space.

We define the semantic attribute direction using endpoint embeddings [see Eq. (8)]:

$$\mathbf{d}_{\text{clip}} = \frac{\mathbf{v}_{\max} - \mathbf{v}_{\min}}{\| \mathbf{v}_{\max} - \mathbf{v}_{\min} \|} \tag{8}$$

where, $\mathbf{v}_{\min}$ and $\mathbf{v}_{\max}$ correspond to images generated at $\alpha = -1$ and $\alpha = +1$, respectively.

To ensure symmetry, we define the semantic midpoint as Eq. (9):

$$\mathbf{v}_{\text{mid}} = \frac{\mathbf{v}_{\max} + \mathbf{v}_{\min}}{2} \tag{9}$$

The directional projection score is then defined as Eq. (10):

$$s_i = (\mathbf{v}_i - \mathbf{v}_{\text{mid}}) \cdot \mathbf{d}_{\text{clip}} \tag{10}$$

This projection extracts only the component aligned with the semantic attribute direction.

1) Why projection instead of distance?

Conventional CLIP evaluation typically relies on global similarity or cosine distance measures. However, distance-based metrics do not distinguish between intended attribute variation and unrelated semantic drift.

Directional projection isolates attribute-consistent change while suppressing orthogonal variation.

2) Monotonic consistency in semantic space: To evaluate whether semantic progression aligns with latent parameter α, we compute the Spearman Rank Correlation [see Eq. (11)]:

$$\rho_{\text{clip}} = \text{Spearman}(\{\alpha_i\}, \{s_i\}) \tag{11}$$

Interpretation:

- $\rho_{\text{clip}} \approx 1 \rightarrow$ monotonic semantic progression.
- $\rho_{\text{clip}} \approx 0 \rightarrow$ semantic incoherence.
- $\rho_{\text{clip}} < 0 \rightarrow$ semantic inversion.

Thus, semantic validation tests whether geometric consistency propagates into the embedding space.

3) Duality with latent QC: The proposed semantic validation mirrors latent qc. The human-layer monotonicity measure is formally defined in Section III-E: ρ_{human}

TABLE I. DUAL STRUCTURAL CRITERIA ACROSS LAYERS

Layer	Structural Criterion
Latent Space	Midpoint symmetry (Δ, c)
Semantic Space	Directional monotonicity (ρ_{clip})
Human Space	Perceptual monotonicity (ρ_{human})

Table I summarizes the structural criteria used at each evaluation layer, where each layer evaluates structural consistency with respect to a distinct representational geometry. This framework enforces geometric validity in latent space, semantic monotonicity in embedding space, and perceptual monotonicity in human evaluation.

E. Human Perceptual Validation

While Section III-C and Section III-D establish geometric and semantic consistency, perceptual reliability ultimately depends on human discernment.

To address this, we introduce a psychophysically grounded evaluation based on pairwise comparison.

1) Two-Alternative Forced Choice (2AFC): For each attribute, we compare generated images along the latent axis using the two-alternative forced choice (2AFC) Protocol.az(α_i)

Given a pair of images (x_i, x_j), participants are asked:

- Which image exhibits a stronger perceptual intensity of attribute a?

2AFC is preferred over absolute rating scales because:

- It reduces inter-observer bias.
- It avoids arbitrary internal calibration.
- It yields statistically consistent pairwise comparison data.

The outcome of the comparison shall be denoted as:

$$x_i \succ x_j$$

if image x_i is judged stronger than x_j.

2) Bradley–Terry preference model: We employ the Bradley–Terry Model to derive a continuous perceptual scale based on pairwise outcomes. Each image is associated with a latent perceptual strength parameter. The probability that is preferred over is defined as: $x_i\theta_i x_i x_j$ [see Eq. (12)]

$$P(x_i \succ x_j) = \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_j)} \quad (12)$$

The parameters θ_i are estimated using maximum likelihood. This model yields a continuous perceptual strength scale:

$$\theta(\alpha_i)$$

associated with each generation parameter α_i.

3) Monotonic consistency in human space: To evaluate Perceptual Monotonicity, we compute the Spearman Rank Correlation between latent parameter and perceptual strength: α [see Eq. (13)]

$$\rho_{\text{human}} = \text{Spearman}(\{\alpha_i\},\{\theta_i\}) \quad (13)$$

Interpretation:

- $\rho_{\text{human}} \approx 1 \rightarrow$ monotonic perceptual progression.
- $\rho_{\text{human}} \approx 0 \rightarrow$ perceptual incoherence.
- $\rho_{\text{human}} < 0 \rightarrow$ perceptual inversion.

Human perceptual validation tests whether geometric and semantic consistency are reflected in subjective discernment. This completes the tri-layer validation structure of the proposed framework.

F. Cross-Layer Consistency Principle

The preceding sections introduce structural validation criteria across three representational layers:

- Latent geometry (Δ, c),
- Semantic embedding space ($\rho_{\text{clip}} > 0$),
- Human perception ($\rho_{\text{human}} > 0$).

These sections consolidate the principles integrating the stated criteria.

1) Necessary and progressive conditions: Geometric validity in latent space is a necessary prerequisite for ensuring perceptual reliability. When midpoint symmetry is violated (indicated by a Large Δ or Low C), the attribute axis becomes structurally distorted. In such cases, semantic or perceptual monotonicity loses meaningful interpretability.

Semantic consistency constitutes an important second-level condition. Even if latent geometry is coherent, the semantic embedding space must maintain directional monotonicity:

$$\rho_{\text{clip}} > 0.$$

Finally, perceptual reliability requires monotonic progression in human discernment:

$$\rho_{\text{human}} > 0.$$

2) Formal definition: We define the perceptual reliability of an attribute as the conjunction of cross-layer monotonicity conditions: a [see Eq. (14)]

$$\text{Reliable}(a) \iff \begin{cases} \Delta \text{ is small}, \\ c \text{ is near } 1, \\ \rho_{\text{clip}} > 0, \\ \rho_{\text{human}} > 0. \end{cases} \quad (14)$$

In practice, threshold values were empirically determined through preliminary validation and are specified in Section IV-B. We determine the thresholds based on preliminary validation of ranges that must be exceeded for geometric distortion to occur.

3) Interpretation: The proposed framework does not assume that controllable generation is universally reliable. Instead, it provides a structured mechanism for diagnosing where reliability fails:

- Latent distortion (geometric failure),
- Semantic misalignment (embedding failure),
- Perceptual ambiguity (human inconsistency).

Thus, perceptual reliability is treated as an emergent property of cross-layer structural consistency.

4) Theoretical implication: The three-layer framework suggests that controllable diffusion generation is not solely determined by latent manipulability. Instead, attribute reliability depends on whether structural coherence propagates across representational geometries:

$$\mathbb{R}^d \rightarrow \mathbb{R}^D \rightarrow \text{Human perceptual space.}$$

This perspective shifts evaluation from measuring similarity to analyzing structural consistency.

IV. Experimental Setup

A. Diffusion Model Configuration

To apply the proposed three-layer framework, we generated images using a Stable Diffusion 1.5–based inpainting model

(Realistic Vision V5.1). We employed an inpainting configuration to enable controlled attribute manipulation while preserving subject identity and overall scene structure.

Unlike full image synthesis, the inpainting setting allows the torso, pose, and background to remain fixed while restricting modifications to the garment region. This configuration isolates attribute variation from confounding factors such as posture, lighting, or composition changes. Accordingly, perceptual variations caused by the latent parameter α are primarily due to garment-related transformations.

Let $z_0 \in \mathbb{R}^d$ denote the latent representation of a base image under a fixed subject condition. For each attribute a, we generated minimal and maximal endpoint latent codes $z_{a,\min}$ and $z_{a,\max}$ using prompt-controlled inpainting. These endpoints define the latent attribute direction introduced in Section III-B.

All latent manipulations occurred in the VAE latent space of the underlying Stable Diffusion architecture. The latent dimensionality is denoted by d, and the CLIP embedding dimensionality is denoted by D, consistent with the definitions in Section III-D.

After geometric validation (Section III-C), we generated α-series images using spherical linear interpolation (SLERP) between endpoints. SLERP preserves latent distribution geometry by interpolating along a hyper-spherical path instead of a straight Euclidean line. This reduces distributional drift during interpolation [17]. To improve perceptual discriminability, we distributed α values using logarithmic scaling and extended their range:

$$\alpha \in [-1.5, 1.5].$$

This extension preserves geometric consistency while increasing perceptual separability in attributes with weaker visual gradients.

Notably, generation and evaluation were structurally separated. The diffusion model serves solely as a transformation mechanism in latent space. We subsequently performed structural validation across three independent representational layers:

- Latent space (geometric QC: Δ, c)
- Semantic embedding space (projection-based monotonicity: ρ_{clip})
- Human perception (2AFC + Bradley–Terry: ρ_{human})

As illustrated in Fig. 1, perceptual reliability is not treated as a property of a single representational space.

Instead, the generation parameter α is propagated across three structurally distinct layers: latent space geometry, semantic embedding space, and human perceptual discernment. Monotonic consistency with respect to α is evaluated independently at each layer. We, therefore, define perceptual reliability as the preservation of cross-layer order rather than isolated similarity within any single space.

Fig. 2 shows the representative base images for each gender–garment condition. In all conditions, the torso, pose, and background remain constant, with only the garment region undergoing subsequent attribute manipulation.

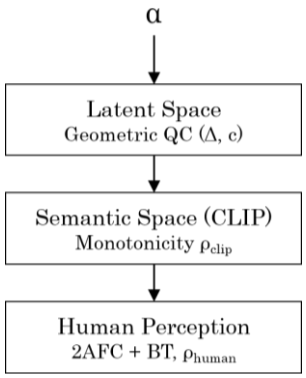


Fig. 1. Three-layer evaluation framework for perceptual reliability.



Fig. 2. Representative base images under fixed gender–garment conditions before attribute manipulation.

B. Latent Direction Construction and Geometric Quality Control

For each perceptual attribute a, latent endpoint representations were constructed under constant subject and garment conditions using the inpainting configuration described in Section IV-A. Using controlled prompt variations, we generated minimal and maximal attribute expressions that yielded the latent codes $z_{a,\min}$ and $z_{a,\max}$ in the VAE latent space of the underlying Stable Diffusion architecture.

We constructed the latent attribute direction $\mathbf{d}_a$ as defined in Eq. (1). This endpoint-based formulation derives the semantic axis in embedding space directly from the generative transformation itself, not from external text embeddings. As a result, the semantic direction remains structurally aligned with the latent manipulation.

Because diffusion latent spaces are high-dimensional and stochastic, simply constructing endpoint representations do not ensure a coherent one-dimensional attribute axis. To ensure structural validity prior to semantic or human evaluation, we employed a geometric validation. For each base image z_0, the

linearity deviation Δ, and centrality $c$were computed as defined in Eq. (5) and Eq. (6). For each attribute under each experimental condition, average values $\bar{\Delta}$ and $\bar{c}$ were obtained across base samples.

We defined the quality control criteria as:

- PASS: $|\bar{\Delta}| < 8.0$ and $\bar{c} > 0.30$
- WEAK-PASS: $|\bar{\Delta}| < 16.0$ and $\bar{c} > 0.15$
- FAIL: otherwise

We excluded attribute directions marked as FAIL and retained WEAK-PASS directions only when visual inspection confirmed they were perceptually interpretable with no visually observable structural distortion.

QC, thereby functions as a structural precondition for cross-layer analysis. If midpoint symmetry is substantially violated, monotonic consistency in semantic embedding space or human perceptual space cannot be meaningfully interpreted. We only propagated attribute directions that satisfied QC to the α-series generation and subsequent cross-layer evaluation.

Fig. 3 shows the representative endpoint.



Fig. 3. Representative minimal (min) and maximal (max) endpoint images for selected attributes.

The upper row corresponds to the fit attribute in the male T-shirt condition, illustrating a geometrically stable latent direction that creates a smooth variation in garment tightness. The lower row shows the pattern scale in the female dress-shirt condition, representing a geometrically weaker (WEAK-PASS) configuration with structurally complex visual variation. These examples offer visual context for the geometric differences evaluated before cross-layer analysis.

C. Semantic Projection Evaluation

To evaluate the preservation of latent attribute progression in the semantic embedding space, we analyzed generated images using a CLIP image encoder.

For each generated image x_i along the α-series, we computed the corresponding image embedding $\mathbf{v}_i \in \mathbb{R}^D$ defined in Eq. (7). Consistent with the endpoint-based formulation introduced in Section III-D, we then constructed the semantic attribute direction using the endpoint embeddings defined in Eq. (8). The semantic midpoint and directional projection scores were computed following Eq. (9) and Eq. (10), respectively.

This projection-based evaluation isolates variation aligned with the semantic attribute direction while suppressing orthogonal changes in embedding. Unlike global similarity or cosine distance measures, directional projection explicitly assesses if attribute progression remains a consistent semantic axis.

We evaluated monotonic consistency in semantic space, using the Spearman rank correlation defined in Eq. (11). A positive correlation indicates that semantic embedding responses preserve the ordering induced by latent manipulation, while near-zero or negative values indicate semantic incoherence or inversion.

Importantly, we treated semantic validation as an intermediate structural layer, not as a substitute for human evaluation. This verifies whether geometric consistency in the latent space propagates to the embedding space before assessing perceptual monotonicity.

D. Human Perceptual Evaluation

To evaluate how latent attribute manipulation reflects human perception, we conducted a two-alternative forced-choice (2AFC) experiment.

We recruited forty participants assigned to experimental conditions defined by gender and garment type. After excluding three responses due to incomplete submissions, 37 valid responses remained for analysis. For each perceptual attribute, participants viewed pairwise comparisons of images generated along the α-series and selected the image with the stronger perceptual intensity.

We chose pairwise comparison over absolute rating scales to reduce inter-observer calibration bias and improve statistical consistency. The 2AFC protocol produces ordinal preference data suitable for probabilistic modeling of latent perceptual strength.

We then estimated perceptual strength parameters θ_i using the Bradley–Terry model defined in Eq. (12). The resulting preference scale provides a continuous representation of perceived attribute intensity for each generation parameter α_i.

To assess perceptual monotonicity, we computed Spearman's rank correlation coefficient between the generation parameter α and the estimated perceptual strength, as defined in Eq. (13). A positive correlation coefficient indicates monotonic perceptual progression, whereas near-zero or negative values indicate perceptual incoherence or inversion.

We randomized image order and left–right presentation to mitigate positional bias. Human evaluation constitutes the final validation layer in the proposed framework. Section V analyzes these results to assess cross-layer alignment. The framework is

designed as a diagnostic tool to identify structurally unreliable attribute manipulations before deployment.

V. Results

Table II summarizes cross-layer monotonicity results aggregated by attribute. All attributes satisfied geometric admissibility, with fit, lightness, and glossiness classified as PASS, and pattern scale categorized as WEAK-PASS according to the criteria specified in Section IV-B.

Semantic monotonicity was consistently positive across all attributes, with high values of ρ_{clip} indicating stable directional progression in embedding space. These results confirm that latent manipulations were structurally preserved at the semantic level.

TABLE II. Cross-Layer Monotonicity Results Aggregated by Attribute.

Attribute	QC Status	ρ_{clip}	ρ_{human}	Cross-layer Alignment
Fit	PASS	0.97	0.67	Strong
Lightness	PASS	0.99	0.31	Moderate
Glossiness	PASS	1.00	0.48	Moderate
Pattern Scale	WEAK-PASS	1.00	0.51	Limited

Perceptual monotonicity, evaluated with 37 valid participant responses, exhibited attribute-dependent variation. Fit demonstrated the strongest perceptual monotonicity. Glossiness exhibited moderate perceptual monotonicity, whereas lightness was comparatively weaker. This suggests that although luminance variation is consistently represented in embedding space, its perception is affected by contextual factors such as shading and local contrast. Pattern scale exhibited limited perceptual reliability despite positive semantic monotonicity.

For fit, latent manipulation primarily affected the garment's contour and silhouette, producing visibly coherent adaptive fitting—tightening and loosening—that appeared across geometric, semantic, and perceptual layers. In contrast, the pattern scale involved changes in motif size together with variations in texture density and contrast. As illustrated in Fig. 3, pattern enlargement introduced composite visual effects beyond simple geometric scaling. This factor presumably reduced perceptual monotonicity despite preserved semantic ordering.

All reported correlations were statistically significant at $p < 0.01$.

Joint evaluation of geometric validity, semantic embedding behavior, and human evaluation shows that perceptual reliability varies across attributes. Fit satisfies all cross-layer conditions defined in Eq. (14), demonstrating strong alignment across representational layers. In contrast, pattern scale demonstrates geometric admissibility but reduced perceptual monotonicity, indicating partial cross-layer misalignment.

These findings support the formulation that perceptual reliability cannot be determined solely from geometric or semantic validity, but arises only when monotonic ordering is preserved across latent geometry, embedding space, and human perception.

VI. Discussion

In this study, we introduced a three-layer evaluation framework to assess perceptual reliability in controllable diffusion-based fashion image generation. By explicitly separating latent geometry, semantic embedding space, and human perception, our proposed framework enables structured diagnosis of attribute controllability across representational levels.

A. Significance of the Three-Layer Decomposition

The central contribution of this study lies in the structural separation of representational layers. Previous evaluations of controllable generation typically rely on either automatic semantic metrics or human preference modeling independently. However, as demonstrated in the present results, geometric validity in latent space does not automatically imply perceptual reliability.

Our proposed framework demonstrates that perceptual reliability is achieved only when the monotonic order induced by the generation parameter α is preserved across all layers. This condition enables identification of alignment failure in geometric construction, semantic embedding behavior, or human perceptual interpretation, thereby providing diagnostic capabilities essential to understanding controllable diffusion models beyond visual realism.

B. Attribute-Dependent Reliability

The results reveal that perceptual reliability is attribute-dependent. Fit exhibited strong cross-layer alignment, satisfying geometric, semantic, and perceptual monotonicity conditions. This suggests that attributes closely tied to structural garment draping are more consistently encoded across representational spaces.

In contrast, pattern scale demonstrated reduced perceptual monotonicity despite geometric admissibility and positive semantic monotonicity. This indicates that certain attributes may be geometrically constructible yet perceptually ambiguous. Such ambiguity may stem from the composite nature of pattern scale, which includes density, contrast, and texture complexity rather than a single structural factor.

Glossiness exhibited moderate behavior, while lightness showed comparatively weaker perceptual monotonicity despite satisfying geometric and semantic criteria. These findings highlight that perceptual interpretation may depend on contextual or experiential factors even when semantic embeddings capture directional consistency.

C. Implications for Automatic Evaluation Metrics

The results further demonstrate that embedding-based monotonicity cannot replace human evaluation. While ρ_{clip} remained positive for all attributes, perceptual monotonicity varied significantly. This partial decoupling suggests that the semantic embedding space captures structural ordering but does not fully encode perceptual salience.

Consequently, evaluation pipelines that rely solely on CLIP similarity or reward models may overestimate controllability when perceptual ambiguity exists. Our proposed cross-layer

framework provides a robust and effective approach to identifying such discrepancies.

D. The Role of Geometric Quality Control

Geometric quality control (QC) plays a foundational role in the framework. By validating midpoint symmetry and collinearity before semantic and perceptual evaluation, QC establishes a critical structural condition for meaningful monotonic analysis.

However, the WEAK-PASS cases indicate that geometric symmetry alone does not determine perceptual reliability. Instead, QC acts as a structural filter that ensures interpretability of subsequent layers, while final reliability depends on cross-layer alignment.

E. Limitations and Future Work

Several research limitations should be noted. First, our study employed only one diffusion backbone and inpainting configuration. Although the proposed framework is model-agnostic, validating it empirically across additional architectures could improve generalizability.

Second, we relied solely on visual discernment for perceptual evaluation. Attributes such as glossiness and fit may also depend on tactile or experiential cues not captured in image-based evaluation. Extending the framework to multimodal perceptual assessment is therefore an important consideration for future research.

Third, we did not explicitly analyze cultural and experiential variability among participants. We recognize that future studies should examine whether perceptual monotonicity differs by demographic or expertise groups.

Finally, several methodological limitations should be acknowledged. The selection of endpoint latent codes $z_{a,min}$ and $z_{a,max}$ was based on controlled generation and visual inspection, and was not formulated as a systematic or optimization-based procedure, which may affect reproducibility. In addition, although the human evaluation employed a two-alternative forced-choice design with 37 participants, a formal statistical power analysis was not conducted, and potential order effects or response dependencies were not explicitly modeled. Furthermore, the study focused on four representative attributes, which may introduce selection bias. Addressing these limitations through principled endpoint selection, more rigorous experimental design, and broader attribute coverage remains an important direction for future work.

F. Broader Implications

The proposed framework could also apply to other fields beyond fashion-image generation, whereby controllable generative models are assessed against human interpretation. Separating geometric validity, semantic embedding behavior, and perceptual monotonicity provides a generalizable framework for assessing human-aligned controllability.

Our study moves from assessing methods that rely solely on isolated similarity to a broader cross-layer monotonicity approach, providing a structured framework for understanding how generative transformations propagate across representational geometries.

VII. Conclusion

In this study, we proposed a three-layer evaluation framework for assessing perceptual reliability in controllable diffusion-based fashion image generation. Explicitly separating latent geometry, semantic embedding space, and human perception allowed for structured validation of attribute manipulation across representational levels.

We introduced geometric quality control to ensure midpoint symmetry and collinearity in the latent space prior to semantic or perceptual evaluation, and evaluated semantic consistency using projection-based monotonicity in the CLIP embedding space. We subsequently assessed perceptual monotonicity using a 2AFC experiment with 37 valid participants, applying the Bradley–Terry model.

Empirical results show that perceptual reliability depends on the attribute being tested: Fit exhibited strong cross-layer alignment, whereas pattern scale showed limited perceptual monotonicity despite geometric admissibility. These findings confirm that controllable generation cannot be evaluated solely through semantic similarity or latent manipulability. Instead, perceptual reliability arises only when a monotonic ordering is preserved across latent geometry, embedding space, and human judgment.

Our proposed framework provides a structured methodology for diagnosing controllable generation beyond visual realism and offers a foundation for human-aligned evaluation of generative models.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach et al., "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 10684–10695, 2022.

[3] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), pp. 8748–8763, 2021.

[4] J. Xu et al., "ImageReward: Learning and evaluating human preferences for text-to-image generation," arXiv preprint arXiv:2304.05977, 2023.

[5] Y. Shen, G. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 9240–9249, 2020.

[6] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "GANSpace: Discovering interpretable GAN controls," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 9841–9850, 2020.

[7] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 3836–3845, 2023.

[8] H. Wang et al., "MV-VTON: Multi-view virtual try-on with diffusion models," arXiv preprint arXiv:2404.17364, 2024.

[9] Z. Zhang, L. Wang, and W. Ding, "LiteMP-VTON: A knowledge-distilled diffusion model for realistic and efficient virtual try-on," Information, vol. 16, no. 5, p. 408, 2025.

[10] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4401–4410, 2019.

[11] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations," in Proc. Int. Conf. Mach. Learn. (ICML), pp. 4114–4124, 2019.

[12] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, 2013.

[13] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it?," IEEE Signal Process. Mag., vol. 26, no. 1, pp. 98–117, 2009.

[14] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 586–595, 2018.

[15] M. Jogan and A. A. Stocker, "A new two-alternative forced choice method," J. Vis., vol. 14, no. 3, 2014.

[16] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs," Biometrika, vol. 39, no. 3–4, pp. 324–345, 1952.

[17] T. White, "Sampling generative networks," arXiv preprint arXiv:1609.04468, 2016.

[18] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023.

[19] J. Ho and T. Salimans, "Classifier-free diffusion guidance," arXiv:2207.12598, 2022.

Performance Comparison of Regularization Methods on Transfer Learning Algorithm for Fish Species Classification

Handrie Noprisson, Anita Ratnasari, Sri Dianing Asri, Vina Ayumi, Hadiguna Setiawan
Faculty of Informatics and Engineering-Department of Informatics, Universitas Dian Nusantara, Indonesia

***Abstract*—Automatic fish classification played an essential role in the fisheries sector, particularly in underwater environments where visual quality was often degraded. This study addressed challenges related to low-contrast underwater images and limited dataset conditions by integrating Contrast Limited Adaptive Histogram Equalization (CLAHE) with a VGG16-based transfer learning model with regularization approaches including L1, L2, and Dropout. The dataset consisted of multiple fish species, including Bream, Sea Bass, Horse Mackerel, Red Mullet, and Black Sea Sprat. To enhance dataset diversity, data augmentation was performed using geometric transformations such as rotation, flipping, cropping/resizing, translation, shearing, and zooming. The dataset was divided into training (70%, 18,900 images), validation (20%, 5,400 images), and testing (10%, 2,700 images). Experimental results showed that the VGG16-CLAHE-Dropout model achieved the best overall performance, with training, validation, and testing accuracies of 99.15%, 98.37%, and 97.07%, respectively. CLAHE was implemented using a clip limit of 2.0 and a tile grid size of 8×8 to enhance image contrast, while the model was optimized using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32. These findings demonstrated that combining contrast enhancement with appropriate regularization techniques significantly improved deep learning performance for underwater fish species classification.**

***Keywords*—Fish classification; underwater image; VGG16; CLAHE; regularization; transfer learning**

I. Introduction

Automatic fish classification is important because it allows the fish industry to be more productive through automation. In addition, the challenges of increasing food demand and potential food scarcity in the future urgently require automatic fish classification [1]. Fish species identification research is important for researchers in various fields such as ecologists, taxonomists, and geneticists to avoid mistaken identification [2]. In addition, automatic fish classification is important in freshwater environments where different species may be similar to each other, making manual identification difficult and prone to error [3]. To increase productivity, support research, ensure accurate species identification, and understand the health benefits of fish is the urgency of automated fish classification research using artificial intelligence [4], [5].

Detection of fish species is important to support fish biodiversity monitoring activities that are used as indicators of the health of seawater ecosystems [6]. Several studies have been conducted to solve the problems that exist in the detection of fish species [7], [8], [9], [10], [11]. In addition, fish species detection research can be useful for calculating the total and type of fish caught as a report to avoid overfishing in a particular area [12].

Unlike other types of data, videos and images cannot be immediately used for automatic detection [13], [14], [15]. The images collected are usually analyzed by manually trained human resources for reporting on the health of marine ecosystems or reporting on fishing areas, which is costly and time-consuming [16]. To overcome this problem, there is a need for an approach in modeling fish species detection technology [17]. This model can be used to help extract information about fish species and their numbers according to the actual situation [18].

Recent studies on fish species classification using deep learning demonstrated notable improvements in performance alongside the evolution of network architectures. ResNet50 and ResNet101 performance remained relatively limited in certain scenarios, with reported accuracies of 91.37% for ResNet50 and 86.12% for ResNet101 [19]. In contrast, MobileNetV2 achieved an accuracy of 96.56% [20], while MobileNetV3 further improved performance to 97.20% [21]. Additionally, integrating image enhancement techniques proved beneficial; for instance, the combination of ESRGAN with VGG16 achieved 96.50% accuracy [22], demonstrating that super-resolution preprocessing significantly improved feature quality in low-resolution images [23].

However, many fish classification techniques rely on transfer learning to improve accuracy. The VGG16 model has been used for fish classification in previous studies. Previous research proposes a framework combining the VGG16 transfer classification method and object localization that proposes rapid classification performance for fish image classification processing [4], [24], [25], [26], [27], [28].

In a study comparing several different transfer learning models, VGG16 achieved the highest accuracy scores, with a testing accuracy of 98.07% without fine-tuning and 96.56% with fine-tuning [20]. However, some previous studies have not discussed the comparative effectiveness of regularization methods applied to VGG16 for fish classification.

Classification of fish species in underwater images is a challenging task due to the phenomenon of illumination in seawater that causes low-contrast imagery. In addition, the scarcity of sufficient species datasets of fish imagery poses a serious problem for training a classification model. To overcome this problem, transfer learning-based methods need to be applied

contrast enhancement methods to improve classification accuracy [18], [29].

The contrast enhancement approach that can be used is Contrast Limited Adaptive Histogram Equalization (CLAHE). By applying the CLAHE method, it is expected to overcome the problem of limited contrast fish species datasets and provide solutions for accurate fish identification. In addition, this study will use the regularization method on VGG16 to reduce overfitting when processing image datasets from contrast enhancement results. In this study, VGG16 was regularized and optimized with the CLAHE method. This proposed research aims to improve the performance of VGG16 in processing species image datasets with low contrast due to lighting conditions in seawater.

II. Research Methodology

The study employed a structured methodology starting with the dataset collection of fish species images, followed by data preprocessing using CLAHE to enhance image quality, and model development using transfer learning based on VGG16 with the application of regularization techniques, including Dropout, L1, and L2. Subsequently, performance evaluation was conducted through benchmarking to compare the effectiveness of each regularization method in improving classification accuracy. In this study, a comparison of different regularizations is proposed to be applied to the VGG16 model layer. The research stage can be seen in Fig. 1.

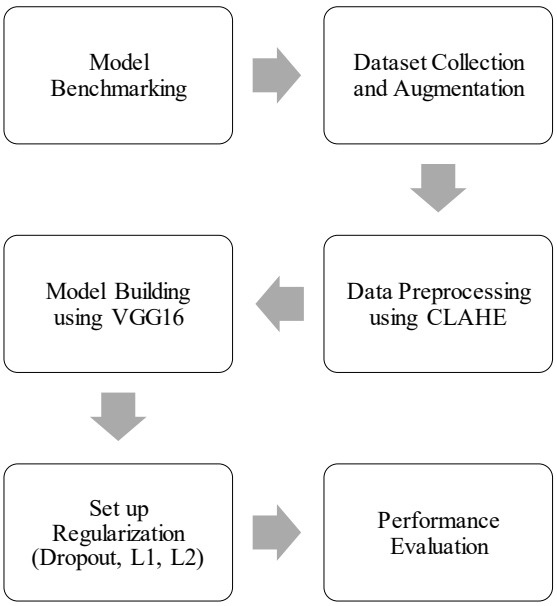


Fig. 1. Research phase

The data classes in this study consisted of Bream, Red Sea Bream, Sea Bass, Red Mullet, Horse Mackerel, Black Sea Sprat, and Striped Red Mullet, while the training and validation datasets were obtained from [30]. Data augmentation was applied to increase dataset diversity and reduce overfitting by performing geometric transformations including rotation, flipping, cropping/resizing, translation, shearing, and zooming. The dataset was divided into three subsets with a composition of 70% (18,900 images) for training, 20% (5,400 images) for validation, and 10% (2,700 images) for testing. The example of dataset can be seen in Fig. 2.

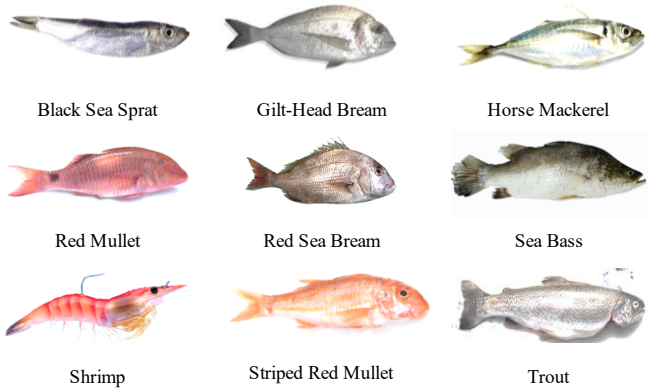


Fig. 2. Fish dataset class

Due to the presence of illumination distortions and low-contrast characteristics typical of underwater imaging, the collected dataset underwent a preprocessing stage employing Contrast Limited Adaptive Histogram Equalization (CLAHE) with a clip limit of 2.0 and a tile grid size of (8×8) to improve image quality. Model construction was performed using the VGG16 architecture with three regularization strategies, i.e., Dropout, L1, and L2, which were implemented in separate experimental configurations. The architecture of VGG16 that is used in this experiment can be seen in Fig. 3.

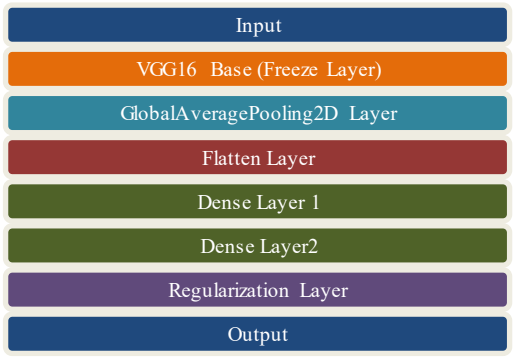


Fig. 3. VGG6 model experiment

Model performance evaluation was conducted using the testing subset through quantitative metrics such as accuracy, confusion matrix, precision, recall, and F1-score.

III. Results and Discussion

The study first conducted a baseline experiment using VGG16 without CLAHE to determine the optimal hyperparameter configuration. Subsequently, CLAHE was applied as a preprocessing technique, and the model was evaluated with different regularization methods. Each configuration was trained and tested under the same experimental settings to ensure a fair comparison.

A. VGG16 Experiment

The experiment evaluated the performance of regularization methods on a transfer learning model based on VGG16 for fish

species classification, with the number of epochs fixed at 50. The learning rate was set to 0.0001, while the optimizer (Adam and SGD) and batch size (16, 32, and 64) were systematically varied to generate six experimental scenarios. The result of experiment is presented in Table I.

TABLE I. VGG16 EXPERIMENT RESULT

No	Learning Rate	Optimizer	Batch Size	Testing Accuracy
1	0.0001	Adam	16	94.85
2	0.0001	Adam	32	95.92
3	0.0001	Adam	64	95.40
4	0.0001	SGD	16	93.75
5	0.0001	SGD	32	94.88
6	0.0001	SGD	64	92.60

The experimental results demonstrated that the choice of optimizer and batch size significantly influenced the performance of the fish classification model. Using the Adam optimizer with a learning rate of 0.0001 consistently produced higher testing accuracy compared to SGD across all batch sizes. The highest accuracy of 95.92% was achieved with a batch size of 32, while batch sizes of 16 and 64 resulted in slightly lower accuracies of 94.85% and 95.40%, respectively. These findings indicated that a moderate batch size provided a better balance between stable gradient updates and generalization capability.

In contrast, the SGD optimizer yielded lower performance overall, with testing accuracies ranging from 92.60% to 94.88%. The best result using SGD was obtained with a batch size of 32 (94.88%), while the lowest accuracy was observed at a batch size of 64 (92.60%). This suggested that SGD was less effective in handling the complexity of underwater fish image classification under the given configuration. Overall, the results confirmed that the Adam optimizer, combined with an appropriate batch size, particularly 32 for fish species classification tasks.

B. VGG16-CLAHE-L1

The accuracy results of the VGG16-CLAHE-L1 model showed that the training process initially started with 0.19 for training and 0.18 for validation, but it quickly decreased and stabilized around 0.10–0.11 for both training and validation in subsequent epochs. Throughout the training process up to 50 epochs, the accuracy values fluctuated slightly but remained relatively constant. The accuracy results for experiments using the VGG16-CLAHE-L1 model can be seen in Fig. 4.

The confusion matrix was used to evaluate the performance of the VGG16-CLAHE-L1 model during the testing phase in classifying fish species across nine categories. The results showed that the model correctly classified certain classes with relatively high accuracy, particularly Horse Mackerel (HM), which achieved 298 correct predictions out of 300 samples. Similarly, Gilt-head Bream (GHB) and Red Mullet (RM) recorded moderate correct classifications with 203 and 112 correctly predicted samples, respectively. However, some classes, such as Black Sea Sprat (BSS) and Sea Bass (SB), were not correctly classified, as all their samples were misclassified, primarily into the Horse Mackerel (HM) class. This indicated that the model had difficulty distinguishing between visually similar species, as presented in Fig. 5.

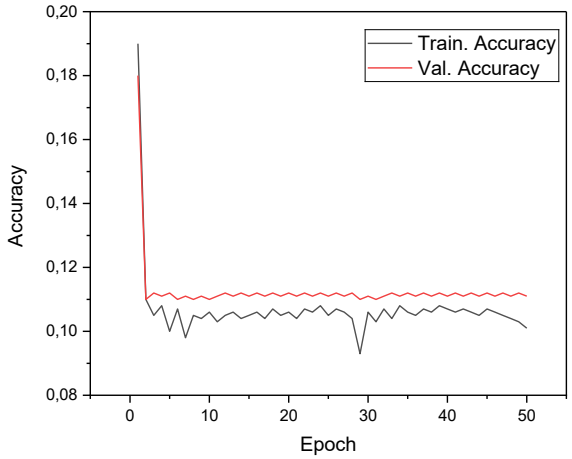


Fig. 4. Accuracy of VGG16-CLAHE-L1 model.

True \ Pred	BSS	GHB	HM	RM	RSB	SB	SH	SRM	TR
BSS	0	0	300	0	0	0	0	0	0
GHB	0	203	97	0	0	0	0	0	0
HM	0	2	298	0	0	0	0	0	0
RM	0	48	140	112	0	0	0	0	0
RSB	0	246	38	0	16	0	0	0	0
SB	0	98	202	0	0	0	0	0	0
SH	0	15	242	5	0	0	38	0	0
SRM	1	19	268	11	0	0	1	0	0
TR	0	141	154	0	0	0	0	0	5

BSS = Black Sea Sprat; GHB = Gilt-head Bream; HM = Horse Mackerel; RM = Red Mullet; RSB = Red Sea Bream; SB = Sea Bass; SH = Shrimp; SRM = Striped Red Mullet; TR = Trout.

Fig. 5. Confusion matrix of VGG16-CLAHE-L1.

Furthermore, substantial misclassification was observed across multiple classes, where a considerable number of samples were incorrectly predicted as Horse Mackerel (HM). For example, Trout (TR), Sea Bass (SB), and Shrimp (SH) were largely misclassified into the HM category, indicating a strong bias of the model toward this dominant class. In addition, Red Sea Bream (RSB) and Striped Red Mullet (SRM) demonstrated low classification performance, with only a small number of correctly identified instances. Based on the confusion matrix, further performance evaluation was conducted using precision, recall, and F1-score metrics, and the results obtained from the VGG16-CLAHE-L1 model were presented in Table II.

The classification performance of the VGG16-CLAHE-L1 model was assessed using precision, recall, and F1-score metrics for each fish species class. The results indicated substantial disparities in performance, where several classes exhibited extremely low or zero effectiveness. Specifically, Black Sea Sprat (BSS), Sea Bass (SB), and Striped Red Mullet (SRM) recorded zero values across all evaluation metrics, demonstrating that these classes were entirely misclassified. Meanwhile, Gilt-head Bream (GHB) and Red Mullet (RM)

showed moderate performance, with F1-scores of 0.38 and 0.52, respectively, indicating partial classification capability.

TABLE II. PRECISION, RECALL, F1-SCORE OF VGG16-CLAHE-L1

Class	Precision	Recall	F1-Score	Support
Black Sea Sprat	0.00	0.00	0.00	300
Gilt-Head Bream	0.26	0.68	0.38	300
Horse Mackerel	0.17	0.99	0.29	300
Red Mullet	0.88	0.37	0.52	300
Red Sea Bream	1.00	0.05	0.10	300
Sea Bass	0.00	0.00	0.00	300
Shrimp	0.97	0.13	0.22	300
Striped Red Mullet	0.00	0.00	0.00	300
Trout	1.00	0.02	0.03	300

In contrast, certain classes displayed imbalanced metric distributions. Red Sea Bream (RSB) and Trout (TR) achieved perfect precision (1.00), yet their very low recall values resulted in poor F1-scores, indicating that only a few instances were correctly identified despite highly accurate predictions. Similarly, Horse Mackerel (HM) achieved a high recall of 0.99 but suffered from low precision (0.17), suggesting that many samples from other classes were incorrectly predicted as HM. Shrimp (SH) also exhibited high precision (0.97) but low recall (0.13), reinforcing the presence of class imbalance and misclassification bias.

C. VGG16-CLAHE-L2

The experimental results of the VGG16-CLAHE-L2 model demonstrated a consistent improvement in accuracy throughout the training process. Initially, both training and validation accuracies were relatively low, but they increased rapidly during the early epochs, indicating that the model effectively learned relevant features from the dataset. As the training progressed, the training accuracy continued to improve gradually, reaching approximately 0.89–0.90, while the validation accuracy fluctuated within the range of 0.70 to 0.76. These trends suggested that the model was able to capture underlying patterns in the data, although some variability in validation performance remained. The detailed accuracy results of the VGG16-CLAHE-L2 experiments were presented in Fig. 6.

Beyond assessing training and validation, the performance model during the testing phase was further evaluated using a confusion matrix. This approach was employed to assess how well the VGG16-CLAHE-L2 model predicted each fish species class correctly or incorrectly across the entire dataset. The confusion matrix results revealed that several classes were classified with high accuracy, such as Horse Mackerel (HM) and Trout (TR), which achieved perfect classification with 300 correct predictions each. Similarly, Red Sea Bream (RSB) and Shrimp (SH) showed strong performance with 298 and 297 correctly classified samples, respectively. Other classes, such as Gilt-head Bream (GHB) and Red Mullet (RM) also demonstrated relatively high correct predictions, indicating improved discriminative capability of the model compared to previous configurations. The detailed results of this evaluation were illustrated in Fig. 7.

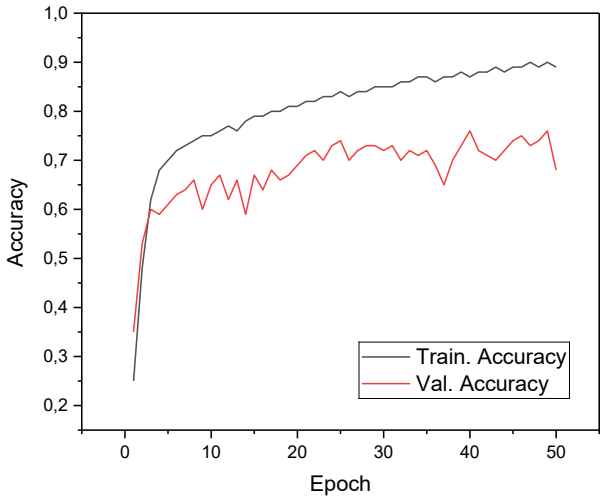


Fig. 6. Accuracy of VGG16-CLAHE-L2.

True \ Pred	BSS	GHB	HM	RM	RSB	SB	SH	SRM	TR
BSS	285	0	2	0	0	1	0	12	0
GHB	0	288	4	0	6	1	0	0	1
HM	0	0	300	0	0	0	0	0	0
RM	0	0	0	243	0	0	0	57	0
RSB	0	2	0	0	298	0	0	0	0
SB	29	5	0	0	0	254	0	0	12
SH	0	0	0	0	0	0	297	3	0
SRM	39	0	24	2	0	0	17	218	0
TR	0	0	0	0	0	0	0	0	300

BSS = Black Sea Sprat; GHB = Gilt-head Bream; HM = Horse Mackerel; RM = Red Mullet; RSB = Red Sea Bream; SB = Sea Bass; SH = Shrimp; SRM = Striped Red Mullet; TR = Trout

Fig. 7. Confusion matrix of VGG16-CLAHE-L2.

Although several classes were classified correctly, a certain level of misclassification was still observed across some categories. Black Sea Sprat (BSS) was occasionally predicted as Striped Red Mullet (SRM), while Sea Bass (SB) showed confusion with both Black Sea Sprat (BSS) and Trout (TR). In addition, Striped Red Mullet (SRM) was misclassified into multiple classes, including BSS and Horse Mackerel (HM), indicating the presence of similar visual features among these species. Despite these limitations, the overall pattern of predictions demonstrated that the VGG16-CLAHE-L2 model achieved a more balanced and reliable performance compared to the L1-based approach. Further evaluation based on the confusion matrix was conducted using precision, recall, and F1-score metrics, where the Red Sea Bream (RSB) class achieved the highest performance across all metrics. The detailed evaluation results for each fish species class were presented in Table III.

The performance of the VGG16-CLAHE-L2 model was evaluated using precision, recall, and F1-score metrics for each fish species class, and the results indicated a substantial improvement compared to the L1-based model. Most classes achieved high and balanced metric values, demonstrating strong

classification capability. Red Sea Bream (RSB) recorded the best overall performance with precision, recall, and F1-score values close to perfect (0.98, 0.99, and 0.99, respectively). Similarly, Trout (TR) and Horse Mackerel (HM) achieved excellent recall values of 1.00, with high precision and F1-scores, indicating that these classes were accurately and consistently classified.

TABLE III. PRECISION, RECALL, F1-SCORE OF VGG16-CLAHE-L2

Class	Precision	Recall	F1-Score	Support
Black Sea Sprat	0.81	0.95	0.87	300
Gilt-Head Bream	0.98	0.96	0.97	300
Horse Mackerel	0.91	1.00	0.95	300
Red Mullet	0.99	0.81	0.89	300
Red Sea Bream	0.98	0.99	0.99	300
Sea Bass	0.90	0.85	0.91	300
Shrimp	0.95	0.99	0.97	300
Striped Red Mullet	0.75	0.73	0.74	300
Trout	0.96	1.00	0.98	300

In addition, Gilt-head Bream (GHB), Shrimp (SH), and Red Mullet (RM) showed strong performance with F1-scores above 0.89, reflecting reliable classification across these categories. Sea Bass (SB) and Black Sea Sprat (BSS) also demonstrated good performance, although slightly lower compared to the top-performing classes. However, Striped Red Mullet (SRM) recorded the lowest performance among all classes, with an F1-score of 0.74, indicating that some misclassification issues persisted.

D. VGG16-CLAHE-Dropout

From the experimental results, the accuracy results of the VGG16-CLAHE-Dropout model showed a strong and consistent improvement throughout the training process. At the initial epochs, both training and validation accuracy started at moderate levels, then increased rapidly as the model learned relevant features from the data. The training accuracy steadily improved and reached approximately 0.94–0.95 in the final epochs, indicating effective learning. The validation accuracy also increased and stabilized in the range of 0.80–0.85, although it exhibited slight fluctuations across epochs. The accuracy results and loss results for experiments using the VGG16-CLAHE-Dropout model can be seen in Fig. 8.

The model performance at the testing stage was further analyzed using a confusion matrix. This evaluation approach was employed to measure how accurately the VGG16-CLAHE-Dropout model classified each fish species across all categories. The results indicated that most classes were correctly predicted with high accuracy, as shown by the dominant diagonal values in the confusion matrix. In particular, Red Sea Bream (RSB) and Shrimp (SH) achieved perfect classification with 300 correct predictions each, while Trout (TR), Gilt-head Bream (GHB), and Sea Bass (SB) also demonstrated near-perfect classification performance, as depicted in Fig. 9.

Nevertheless, a small number of misclassifications were still observed in certain classes. For example, Striped Red Mullet (SRM) was occasionally misclassified into classes such as Black Sea Sprat (BSS) and Horse Mackerel (HM), while minor confusion occurred between Horse Mackerel (HM) and Red Sea Bream (RSB). Despite these errors, the overall classification performance remained highly reliable and balanced across all classes. The next evaluation was conducted using precision, recall, and F1-score metrics. The detailed results for each fish species class were presented in Table IV.

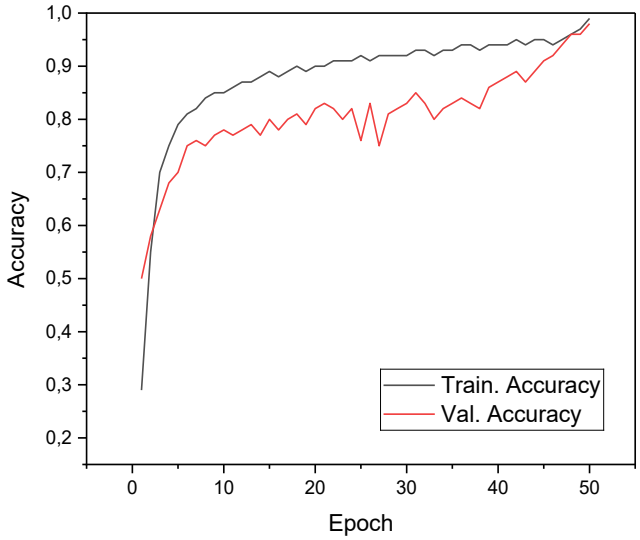


Fig. 8. Accuracy of VGG16-CLAHE-Dropout.

True \ Pred	BSS	GHB	HM	RM	RSB	SB	SH	SRM	TR
BSS	295	0	1	0	0	3	0	1	0
GHB	0	294	0	0	6	0	0	0	0
HM	0	4	288	0	8	0	0	0	0
RM	0	0	0	285	0	0	0	15	0
RSB	0	0	0	0	300	0	0	0	0
SB	5	0	0	0	2	292	0	1	0
SH	0	0	0	0	0	0	300	0	0
SRM	12	0	10	2	1	1	6	268	0
TR	0	0	0	0	0	1	0	0	299

BSS = Black Sea Sprat; GHB = Gilt-head Bream; HM = Horse Mackerel; RM = Red Mullet; RSB = Red Sea Bream; SB = Sea Bass; SH = Shrimp; SRM = Striped Red Mullet; TR = Trout

Fig. 9. Confusion matrix of VGG16-CLAHE-Dropout.

TABLE IV. PRECISION, RECALL, F1-SCORE OF VGG16-CLAHE-DROPOUT

Class	Precision	Recall	F1-score
Black Sea Sprat	0.95	0.98	0.96
Gilt-Head Bream	0.99	0.98	0.98
Horse Mackerel	0.96	0.96	0.96
Red Mullet	0.99	0.95	0.97
Red Sea Bream	0.95	1.00	0.97
Sea Bass	0.98	0.97	0.98
Shrimp	0.98	1.00	0.99
Striped Red Mullet	0.94	0.89	0.92
Trout	1.00	1.00	1.00

Most classes achieved excellent and well-balanced metric values, indicating strong generalization capability of the model. Trout (TR) recorded perfect performance with precision, recall, and F1-score all reaching 1.00, while Shrimp (SH) and Red Sea Bream (RSB) also achieved near-perfect results, with recall values of 1.00 and high precision and F1-scores. Additionally, Gilt-head Bream (GHB), Sea Bass (SB), and Red Mullet (RM) showed strong and stable performance, with F1-scores ranging from 0.97 to 0.98. Furthermore, Black Sea Sprat (BSS) and Horse Mackerel (HM) demonstrated high classification accuracy with balanced precision and recall values around 0.95–0.98. Although Striped Red Mullet (SRM) recorded the lowest performance among all classes, with an F1-score of 0.92, its results remained relatively strong compared to previous models.

E. Discussion

Previous studies proposed various frameworks for fish image classification, including approaches that combined VGG16-based transfer learning to achieve accurate classification performance. Other studies compared multiple transfer learning architectures. However, most of these works did not explore the comparative effectiveness of different regularization techniques applied specifically to the VGG16 architecture. In this study, several configurations of VGG16 were evaluated, including the baseline model and variants integrated with CLAHE and different regularization methods, and the experimental results were summarized in Table V.

TABLE V. COMPARISON OF RESULTS WITH PREVIOUS STUDIES

Source	Model	Testing Accuracy
This research	VGG16	95.92
This research	VGG16-CLAHE-L1	24.89
This research	VGG16-CLAHE-L2	91.96
This research	VGG16-CLAHE-Dropout	97.07
[19]	ResNet50	91.37
[19]	ResNet101	86.12
[20]	MobileNetV2	96.56
[22]	ESRGAN-VGG16	96.50
[21]	MobileNetV3	97.20

The experimental results indicated that the proposed VGG16-CLAHE-Dropout model achieved the highest testing accuracy of 97.07%, outperforming the baseline VGG16 (95.92%) as well as other variants, including VGG16-CLAHE-L2 (91.96%) and VGG16-CLAHE-L1 (24.89%). When compared with prior studies, the proposed approach demonstrated competitive performance, closely approaching MobileNetV3 (97.20%) and surpassing MobileNetV2 (96.56%), ESRGAN-VGG16 (96.50%), and ResNet-based models such as ResNet50 (91.37%) and ResNet101 (86.12%). These findings suggested that the integration of CLAHE and Dropout regularization within the VGG16 framework significantly enhanced classification accuracy and robustness. Furthermore, the architecture of the proposed VGG16-based framework was presented in Table VI.

The proposed framework for fish classification was designed based on a VGG16-based architecture with additional layers for feature refinement and regularization. The model took input images of size 224 × 224 × 3, which were processed through the pre-trained VGG16 convolutional base to extract high-level feature representations. The convolutional layers were frozen to preserve the learned features, followed by a GlobalAveragePooling2D layer that reduced the spatial dimensions into a compact feature vector. Subsequently, a flatten layer was applied to prepare the features for the fully connected layers. The classification stage consisted of two dense layers with 4096 and 1072 neurons, respectively, which enhanced the model's ability to learn complex patterns. A dropout layer was incorporated to reduce overfitting by randomly deactivating neurons during training.

TABLE VI. PROPOSED FRAMEWORK FOR FISH CLASSIFICATION

No	Layer	Layer Type	Output Shape
1	Input	Input Layer	(224, 224, 3)
2	VGG16 Base	Convolutional Blocks	(7, 7, 512)
3	Freeze Layers	-	-
4	GlobalAveragePooling2D	Pooling	512
5	Flatten	Flatten	512
6	Dense 1	Fully Connected	4096
7	Dense 2	Fully Connected	1072
8	Dropout	Regularization	1072
9	Output	Dense	9

IV. CONCLUSION

In conclusion, the training process was conducted to optimize the VGG16 model by minimizing the discrepancy between predicted outputs and ground truth labels, while incorporating a comparative analysis of regularization techniques within the network. The experimental results demonstrated that the VGG16-CLAHE-Dropout model achieved the highest testing accuracy of 97.07%, outperforming the baseline VGG16 (95.92%), VGG16-CLAHE-L2 (91.96%), and VGG16-CLAHE-L1 (24.89%). The model was trained for 50 epochs, and CLAHE was applied using a clip limit of 2.0 and a tile grid size of 8 × 8 to enhance image contrast. Furthermore, the optimization process was carried out using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32.

Future research will focus on improving the robustness and generalization capability of the VGG16-CLAHE framework by exploring more advanced regularization strategies, as well as utilizing larger and more diverse underwater datasets. Additional contrast enhancement techniques tailored for complex aquatic environments will be investigated to further improve feature representation. Moreover, subsequent studies will integrate modern architectures such as EfficientNet, Vision Transformers, and hybrid CNN–Transformer models to evaluate whether they will outperform VGG16 in handling low-contrast fish images.

ACKNOWLEDGMENT

This research is fully supported by the Universitas Dian Nusantara and Direktorat Riset, Teknologi, dan Pengabdian

Kepada Masyarakat, Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia through research funding with No.792/LL3/AL.04/2024 c.q. No.11/136/H-SPK/VI/2024.

REFERENCES

[1] A. Fitzgerald, C. C. Ioannou, S. Consuegra, A. Dowsey, and C. García de Leaniz, "Machine vision applications for welfare monitoring in aquaculture: challenges and opportunities," Aquac. Fish Fish., vol. 5, no. 1, p. e70036, 2025.

[2] M. Sueker et al., "FISH-SPEC: Fast Identification System for Handheld Spectroscopy and Species Classification," Appl. Food Res., p. 101536, 2025.

[3] Z. Farooq, M. Ramzan, M. Bilal, M. Attique, T.-S. Chung, and A. Naz, "A multi-class framework for fish species classification using deep learning technique," PLoS One, vol. 21, no. 2, p. e0342901, 2026.

[4] A. Kuswantori, T. Suesut, W. Tangsrirat, G. Schleining, and N. Nunak, "Fish Detection and Classification for Automatic Sorting System with an Optimized YOLO Algorithm," Appl. Sci., vol. 13, no. 6, p. 3812, 2023, doi: 10.3390/app13063812.

[5] M. Bhanumathi and B. Arthi, "Future Trends and Short-Review on Fish Species Classification Models Based on Deep Learning Approaches," in 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), IEEE, Nov. 2022, pp. 01–05. doi: 10.1109/ICAISS55157.2022.10011087.

[6] M. Pinna et al., "An overview of ecological indicators of fish to evaluate the anthropogenic pressures in aquatic ecosystems: from traditional to innovative DNA-based approaches," Water, vol. 15, no. 5, p. 949, 2023.

[7] C. Yu, Z. Hu, B. Han, Y. Dai, Y. Zhao, and Y. Deng, "An intelligent measurement scheme for basic characters of fish in smart aquaculture," Comput. Electron. Agric., vol. 204, p. 107506, 2023.

[8] Y. Liu et al., "DP-FishNet: Dual-path Pyramid Vision Transformer-based underwater fish detection network," Expert Syst. Appl., vol. 238, p. 122018, 2024.

[9] W. Ji, J. Peng, B. Xu, and T. Zhang, "Real-time detection of underwater river crab based on multi-scale pyramid fusion image enhancement and MobileCenterNet model," Comput. Electron. Agric., vol. 204, p. 107522, 2023.

[10] K. Liu, L. Peng, and S. Tang, "Underwater object detection using TC-YOLO with attention mechanisms," Sensors, vol. 23, no. 5, p. 2567, 2023.

[11] B. Nur Korkmaz, R. Diamant, G. Danino, and A. Testolin, "Automated detection of dolphin whistles with convolutional networks and transfer learning," Front. Artif. Intell., vol. 6, p. 1099022, 2023.

[12] Y. A. Osman and M. Samy-Kamal, "Diversity and characteristics of commercial Red Sea fish species based on fish market survey: informing management to reduce the risk of overfishing," J. Fish Biol., vol. 102, no. 4, pp. 936–951, 2023.

[13] V. Ayumi et al., "Transfer Learning for Medicinal Plant Leaves Recognition: A Comparison with and without a Fine-Tuning Strategy," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 9, 2022, doi: 10.14569/IJACSA.2022.0130916.

[14] H. Noprisson, E. Ermatita, A. Abdiansah, V. Ayumi, M. Purba, and M. Utami, "Hand-Woven Fabric Motif Recognition Methods: A Systematic Literature Review," in 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS, IEEE, 2021, pp. 90–95.

[15] M. Purba et al., "Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 9, 2022, doi: 10.14569/IJACSA.2022.0130917.

[16] E. Varghese, K. R. Sreenath, J. Jayasankar, and V. V. R. Suresh, "Artificial Intelligence for the Blue Revolution: Advancing Fisheries and Aquaculture Management," in Harvesting Intelligence: The AI Revolution in Agriculture: From Fields to Algorithms, Cultivating Future Harvests, Springer, 2026, pp. 197–216.

[17] L. Pawar et al., "Pros and cons of diagnostic methods used for AMR surveillance in aquaculture," in Antimicrobial Resistance in Aquaculture and Aquatic Environments, Springer, 2025, pp. 137–161.

[18] M. Francescangeli et al., "Image dataset for benchmarking automated fish detection and classification algorithms," Sci. data, vol. 10, no. 1, p. 5, 2023.

[19] V. H. Kaya, İ. Akgül, and Ö. Z. TANIR, "IsVoNet8: A Proposed Deep Learning Model for Classification of Some Fish Species," Tarim Bilim. Dergisi-journal Agric. Sci., 2022, doi: 10.15832/ankutbd.1031130.

[20] D. Fitrianah, K. M. Suryaningrum, N. T. M. Sagala, V. Ayumi, and S. M. Lim, "Fine-Tuned MobileNetV2 and VGG16 Algorithm for Fish Image Classification," in 2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), IEEE, 2022, pp. 384–389.

[21] B. Li, Y. Liu, and Q. Duan, "T-KD: two-tier knowledge distillation for a lightweight underwater fish species classification model," Aquac. Int., vol. 32, no. 3, pp. 3107–3128, 2024.

[22] S. Adhikary, S. Banerjee, R. Singh, and A. D. Dwivedi, "Fish species identification on low resolution—a study with enhanced super-resolution generative adversarial network (ESRGAN), YOLO and VGG-16," PeerJ Comput. Sci., vol. 11, p. e2860, 2025.

[23] W. Ilham and A. Ahmad, "A comprehensive review of convnext architecture in image classification: Performance, applications, and prospects," IJACI Int. J. Adv. Comput. Informatics, vol. 2, no. 2, pp. 108–114, 2026.

[24] F. J. P. Montalbo and A. A. Hernandez, "Classification of Fish Species with Augmented Data using Deep Convolutional Neural Network," in 2019 IEEE 9th International Conference on System Engineering and Technology (ICSET), IEEE, Oct. 2019, pp. 396–401. doi: 10.1109/ICSEngT.2019.8906433.

[25] F. Stephenson, J. R. Leathwick, M. P. Francis, and C. J. Lundquist, "A New Zealand demersal fish classification using Gradient Forest models," New Zeal. J. Mar. Freshw. Res., vol. 54, no. 1, pp. 60–85, 2020.

[26] B. S. Rekha, G. N. Srinivasan, S. K. Reddy, D. Kakwani, and N. Bhattad, "Fish detection and classification using convolutional neural networks," in International Conference On Computational Vision and Bio Inspired Computing, Springer, 2019, pp. 1221–1231.

[27] M. K. Alsmadi and I. Almarashdeh, "A survey on fish classification techniques," J. King Saud Univ. Inf. Sci., 2020.

[28] J.-H. Park and Y.-K. Choi, "Efficient Data Acquisition and CNN Design for Fish Species Classification in Inland Waters," J. Inf. Commun. Converg. Eng., vol. 18, no. 2, pp. 106–114, 2020.

[29] M. A. Ahmed, M. S. Hossain, W. Rahman, A. H. Uddin, and M. T. Islam, "An advanced Bangladeshi local fish classification system based on the combination of deep learning and the internet of things (IoT)," J. Agric. Food Res., vol. 14, p. 100663, 2023.

[30] O. Ulucan, D. Karakaya, and M. Turkan, "A Large-Scale Dataset for Fish Segmentation and Classification," in 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE, 2020, pp. 1–5.

FEM-KP: A Functional Evaluation Metric for Keyphrase Prediction Models

Lahbib Ajallouda[1], Ahmed Zellou[2]
LARSLACH Laboratory Poly-Disciplinary Faculty of Es-Semara, Ibn Zohr University, Agadir, Morocco[1]
SPM Research Team ENSIAS, Mohammed V University, Rabat, Morocco[2]

***Abstract*—Keyphrase prediction models are among the natural language processing (NLP) tasks that have improved their performance with transformers and large language models (LLMs). Instead of extracting present keyphrases in the text, these models also generate absent keyphrases. This improvement has led to significant challenges in the evaluation process of these models, which rely on metrics that compare the predicted keyphrases with the reference keyphrases. To measure the performance of these models, several evaluation metrics such as F1-score, ROUGE-L, and BertScore have been used. However, they often prioritize lexical similarity over semantic usefulness. Consequently, the functional usefulness of keyphrases in document representation is not evaluated during the evaluation process, which leads to inconsistencies in the evaluation results. Therefore, in this paper we propose a functional evaluation metric for Keyphrase prediction models (FEM-KP), a new evaluation metric that uses a two-track approach where track (A) evaluates the performance of the model to generate keyphrases capable of constructing a document summary, while track (B) measures the ability of these phrases to retrieve the document. We evaluated the performance of four keyphrase prediction models using current evaluation metrics and FEM-KP across the Inspec, KP20k, and Krapivin datasets. The experimental results showed that FEM-KP is the only evaluation system that maintained a consistent performance ranking regardless of document length or dataset complexity. In contrast, other metrics showed inversions in ranking. These results confirm that FEM-KP is a robust, reliable, and domain-independent evaluation metric for evaluating the performance of keyphrase prediction systems.**

Keywords*—Document retrieval; Document summarization; Evaluation metrics; Functional evaluation metric; Keyphrase prediction models; natural language processing

I. Introduction

Keyphrases prediction is one of the most important tasks in natural language processing [1]. These phrases provide descriptive and semantic information that summarizes the content of documents. Recent studies have demonstrated that exploiting keyphrases improves the performance of summarization models by directing them toward the important topics of the text [2]. Also, the exploitation of keyphrases in the retrieval task improves the quality of retrieved snippets and reduces computational costs [3]. Keyphrases also improve the accuracy of responses in question-and-answer tasks [4]. Furthermore, digital indexation has been improved by using keyphrase generation [5].

Keyphrase prediction models (KPM) are typically evaluated by comparing predicted with referenced keyphrases, without considering semantic aspects [6]. This evaluation does not accurately reflect the performance of these models (KPEVAL). Generative models are more affected by this problem because they generate the keyphrases rather than extracting them from the text [7].

To improve the quality of KPM evaluation, some studies, such as [8], have proposed a framework exploit edit distance, token matching, and duplication penalty. Also, at [6], a two-way comparison was conducted between predicted and reference keyphrases using the BertScore [9]. In contrast, the KPEVAL framework propose to combine accuracy, variety, reliability, and reference consistency. However, several challenges still limit the reliability of the results of these models, such as the reliance on a set of reference keyphrases to compare KPM performance that are often self-defined and influenced by linguistic style. Also, the difficulty to evaluate the functional performance of the keyphrases generated in the document representation.

To overcome these challenges, we propose the FEM-KP metric, a comprehensive metric to evaluate KPM by measuring the functional utility of these phrases rather than linguistic or semantic similarity. The importance of FEM-KP lies in transforming the evaluation process, which focuses on linguistic or semantic similarity between the predicted and reference keyphrases, into a deeper semantic evaluation of the functional quality of these phrases. Our metric adopts two tracks, the first evaluates the ability of the predected keyphrases to generate a document summary, while the second evaluates their ability to retrieve the document.

The content of paper includes, a review of previous studies that focused on keyphrase prediction models, text summarization, text retrieval, and sentence embedding techniques in section 2. The current evaluation methods for keyphrase prediction models are presented in section 3, including evaluation metrics, evaluation datasets, and challenges that limit their reliability. Section 4 provides the proposed FEM-KP metric, its design, and formulation. The experimental results of FEM-KP performance are presented in Section 5. The paper concludes with section 6, which summarizes our work and identifies future research directions.

II. Related Work

To present the mechanisms and techniques that will be exploited in FEM-KP metric, we have devoted this section to review the keyphrase prediction models, sentence embedding techniques, text retrieval techniques, and text summarization techniques.

A. Keyphrase Prediction Models

Keyphrases are expressions consisting of one or more words, that serve not only to index but also as semantic summaries. Keyphrases are exploited in classification, clustering, text summarization, and information retrieval [10]. Keyphrase models have evolved from extractive models based on statistical techniques to generative models based on transformers and large language models (LLMs) [11]. Several reviews have addressed the development of these models, while the review in [12] is focused on statistical and graphical methods, [13] reviewed the improved performance using deep learning to generate keyphrases.

In contrast, other studies, such as [14], focused on reviewing the latest methods that use pre-trained models such as SBERT and RoberTa for keyphrase extraction, and how these models improved extraction results compared to previous methods. In [15], the comparison between extraction and generation mechanism, and the advantages and disadvantages of statistical methods and deep learning methods to extract and generate keyphrases were identified. Furthermore, [16] demonstrated how sentence embeddings can be used to improve contextual extraction performance. Also, [17] reviewed the integration of extraction and generation mechanisms in keyword prediction, proposing a classification of models based on pre-trained models, as well as semi-supervised and unsupervised approaches. In contrast, [18] analyzes the performance of LLM models to extract keyphrases and presents a comparison between Llama3, GPT-4o and other large language models.

B. Text Semantic Representation

With their crucial role in many NLP tasks, text embedding models have garnered increasing attention from researchers, particularly with the emergence of Large Language Models (LLMs) [19]. Text embedding began with models that represented each word individually, such as Word2Vec, which learned word representations from their contexts. Later, GloVe (Global Vectors) exploit also comprehensive word frequency statistics to improve the performance of embedding. However, the challenge of representing words according to their meaning within a sentence remained until the emergence of ELMo as a model based on bidirectional LSTM networks to produce contextual representations based on word meaning within a sentence.

The advent of transformers revolutionized text embedding, with the emergence of the BERT (Bidirectional Coder Representations of Transformers) model, which uses transformer architecture to create deep, and contextual embeddings of words and sentences. This model was further enhanced at the sentence and paragraph by the SBERT model. Multilingual models like LaBSE, which produces comparable embeddings across different languages, have also been developed, proving very useful for multilingual retrieval. Furthermore, the challenge of long documents has led to the emergence of models that represent long documents, such as Dense Passage Retrieval (DPR), which uses BERT-like models to embed long documents. ColBERTv2 also produces efficient embeds by layering the representation to enhance semantic differentiation. In contrast, the emergence of Large Language Models (LLMs) such as GPT4 and LLaMA has contributed to the production of text embeds that exhibit a greater ability to capture semantic relationships, both in retrieval and text generation. OpenAI Embeddings relies on LLM models to generate vectors used in semantic retrieval, Q&A, and classification.

The open-source LLaMA Embeddings is also used to create embeds with greater context and a high capacity for representing long texts. Fig. 1 presents the evolution of text embedding models.

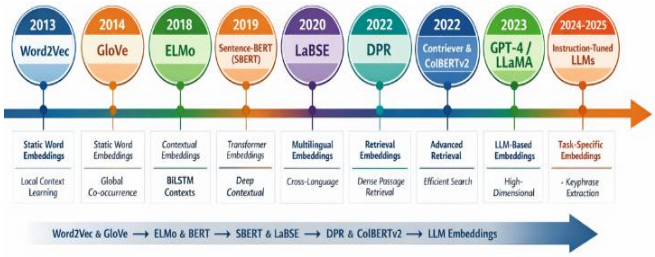


Fig. 1. Evolution of text embedding models.

C. Text Summarization Models

Text summarization tasks have evolved through several approaches. Extractive summarization, which relies on features such as word frequency, position, and similarity metrics, was the first method used text summarization [20]. Graph algorithms such as TextRank and LexRank [21] were also exploited to model the relationships between sentences. With the advent of neural models, particularly RNN/LSTM, abstractive summarization emerged, which focused on paraphrasing the text rather than extracting [22], [23]. The ability of transformers, such as BERT [24] and BART [25], to provide deep contextual understanding and generate texts has also contributed to improving the text summarization task. Large language models (LLMs) such as GPT-3 and GPT-4 have also enhanced text summarization performance [26]. Several reviews have focused on the text summarization task. The impact of large language models on summary coherence in zero-based training environments was studied in [27]. The use of LLMs in the specialize contexts, such as the medical or scientific fields, was also discussed in the studies [28] and [29]. These studies confirmed the significant impact of LLMs on text summarization task.

D. Text Retrieval Models

Text retrieval models are classified into several categories. The first is lexical and statistical matching, which relies on term-frequency-inverse document frequency (TF-IDF) to sort documents, such as BM25 [30], which sorts text based on keyphrases. There is also the category of dense retrieval, which uses transforms to represent text in a vector space, such as the

DPR (Dense Passage Retrieval) model [31] , which exploits bi-encoder to represent the query and the document separately. Also, a hybrid and multi-vector late interaction models, which combine the efficiency of dense retrieval models with the accuracy of lexical retrieval, such as the ColBERT (Contextualized Late Interaction over BERT) model [32], which is more accurate in capturing fine-grained textual details.

Large language models have also contributed to improved retrieval task. Large Language Model-Augmented Retrieval (LLM-Augmented Retrieval), which focuses on using LLMs to generate hypothetical documents or optimize queries before the retrieval process, such as HyDE (Hypothetical Document Embeddings) model [33]. This model generates a hypothetical text from the query and then searches for similar documents. In general, studies indicate that retrieval applications rely on pre-trained embeddings and neural structures to provide more accurate semantic matching [34] and [35], which supports the hypothesis that a keyphrase is a condensed summary from which the original document can be retrieved and reconstructed.

Word embedding models (Word2Vec, GloVe) are considered the first step towards text embedding, followed by the emergence of contextual models (ELMo, BERT), which evolved into sentence and paragraph embedding models (SBERT, LaBSE), then expanded further to contextual retrieval models (DPR, ColBERTv2), and finally to LLM-based embeddings, which are related to the higher text representation capabilities of multiple NLP tasks.

III. Keyphrase Evaluation Paradigms

In recent years, most keyphrase prediction models have adopted a generation mechanism, which has complicated the task of evaluating these models. Most current evaluation metrics rely on lexical matching, which lacks the ability to measure semantic meaning. In this context, this section will provide a review of current evaluation metrics, from statistical to embedding metrics, to highlight the research gap represented by the absence of a standardized metric that combines the precision with the contextual and functional significance of predicted keyphrase.

A. Lexical Evaluation Metrics

Lexical metrics are the most commonly used to evaluate keyphrase prediction models. These metrics calculate the exact match between the keyphrases predicted by the model (KP) and the reference keyphrases (RP).

1) Precision, recall, and f1-measure metrics: Precision, Recall, and F1-measure are used to evaluate keyphrase prediction models based on exact matching. Precision measures the accuracy of predictions, by calculating the proportion of correct phrases extracted from the total identified by the model. It also indicates the model's ability to exclude irrelevant phrases. Recall measures the coverage of reference phrases, by calculating the model's ability to retrieve all reference keyphrases. F1-measure measures the harmonic mean between precision and recall, to achieve a balance between the coverage and the accuracy. Table I presents the mathematical definitions of Precision, Recall, and F1-measure metrics.

TABLE I. Mathematical Formula of Precision, Recall, and F1-Measure Metrics

Metric	Mathematical Formula
Precision	$\frac{\text{correctly predicted keyphrases}}{\text{all predicted keyphrases}}$ (1)
Recall	$\frac{\text{correctly predicted keyphrases}}{\text{all reference keyphrases}}$ (2)
F1-measure	$2 \times \frac{Precision \times Recall}{Precision + Recall}$ (3)

Although the precision, recall, and F1-Measure metrics provide a lexical evaluation of key phrase prediction performance, they rely on exact matching between the predicted and reference keyphrases. However, lexical interference at the n-gram and sequences, which results in partial matching between phrases, is the most significant challenge for these metrics.

2) ROUGE metric: The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric, primarily used to evaluate text summarization models. This metric has been adapted to evaluate keyphrase prediction by treating each predicted keyphrase and reference keyphrase as a sequence of tokens. ROUGE calculates a lexical overlap between predicted keyphrases and reference keyphrases. REUGE is used in various forms. The most commonly used variants are:

ROUGE-N: measures the overlap of n-grams. For example, ROUGE-1 evaluates unigram overlap, and ROUGE-2 evaluates bigram overlap. ROUGE-N is calculated using formula 4.

$$ROUGE - N = \frac{\sum_{gram_n \in Ref} min\,(Count_{pred}(gram_n), Count_{ref}(gram_n))}{\sum_{gram_n \in Ref} Count_{ref}(gram_n))} \quad (4)$$

where:

- Ref: Set of reference keyphrases.
- Pred: set of predicted keyphrases.
- $gram_n$: n-gram token sequence.
- $Count_{pred}(.)$: Occurrences of the n-gram in predicted keyphrases.
- $Count_{ref}(.)$: Occurrences of the n-gram in reference keyphrases.

ROUGE-L: Unlike ROUGE-N, ROUGE-L relies to find the Longest Common Subsequence (LCS) that appear in both the predicted keyphrase and the reference keyphrase. ROUGE-L is calculated using LCS-based decision and recall metrics by formulas 5, 6 and 7.

$$Precision_{LCS} = \frac{LCS(PK,RK)}{|PK|} \quad (5)$$

$$Recall_{LCS} = \frac{LCS(PK,RK)}{|RK|} \quad (6)$$

$$ROUGE - L_{LCS} = \frac{(1+\beta^2) \times Precision_{LCS} \times Recall_{LCS}}{Recall_{LCS} + \beta^2 \times Precision_{LCS}} \quad (7)$$

where:

- RK : reference keyphrases
- PK : predicted keyphrases
- LCS(PK, RK): length of the longest common subsequence between PK and RK.
- |RK| : total number of tokens in reference keyphrases
- |PK| : total number of tokens in predicted keyphrases
- β :A parameter used to set the priority between precision and recall (typically β=1).

While lexical metrics facilitate the evaluation process, they disregard the expected order and importance of keyphrases in the document. This limitation has spurred the use rank-based metrics such as MRR, NDCG, and MAP, which consider the position and importance of the keyphrase.

B. Rank-Based Evaluation Metrics

To overcome the challenges of lexical evaluation, rank-based metrics consider the rinking of predicted keyphrases, where top-ranked candidates are of primary interest. Mean Reciprocal Rank (MRR) is a simple metric that focuses on the rank of the first keyphrase predicted $rank_i$ across a set of N documents. MRR is calculated by formula 8.

$$MRR = \frac{1}{|N|}\sum_{i=1}^{|N|}\frac{1}{rank_i} \quad (8)$$

MRR ignores the rank of other keyphrases is a challenge that limits its ability to evaluate multiple keyphrases. To address this challenge, Mean Average Precision (MAP) metric calculates the average precision for each position where keyphrase occurs. Formula 9 is used to calculate the MAP metric.

$$MAP = \frac{1}{|N|}\sum_{i=1}^{|N|}\left(\frac{1}{|K_i|}\sum_{j=1}^{m}P_i(j)\times rel_i(j)\right) \quad (9)$$

where K_i is the set of reference keyphrases, m is the number of predicted key phrases, $P_i(j)$ denotes precision at rank k, and $rel_i(j)$ is a relevance indicator, is 1 if the predicted keyphrase at rank k is relevant, and 0 otherwise.

MAP metric treats all keyphrases with the same weight, regardless of their importance in the document. To overcome this challenge, the Normalized Discounted Cumulative Gain (NDCG) metric, uses the logarithmic discount, where the importance of the keyphrase decreases as its position in the list decreases. Formula 10 is used to calculate the NDCG metric.

$$\mathrm{NDCG_k} = \frac{\sum_{i=1}^{k}\frac{2^{\mathrm{rel_{i-1}}}}{\log_2(i+1)}}{\sum_{i=1}^{k}\frac{2^{\mathrm{Irel_{i-1}}}}{\log_2(i+1)}} \quad (10)$$

where:

- k: The number of top predicted keyphrases.
- i: The rank position in the predicted list.
- rel_i: The relevance score of the keyphrase at rank i in the predicted ranking.
- $Irel_i$: the relevance score of the keyphrase at rank i in the ideal ranking.
- $log_2(i+1)$: logarithmic discount factor.

Lexical and rank-based evaluation metrics often overlook the semantic comparison of keyphrases. This challenge becomes even more pronounced with absent keyphrases, which may be more semantically accurate but differ literally from the reference keyphrases. Therefore, the study of semantic metrics is essential.

C. Semantic Similarity Metrics

Semantic similarity metrics represent the second generation of evaluation metrics, developed to overcome the limitations of lexical and rank-based metrics. Semantic similarity metrics represente predicted and reference keyphrases in a semantic space using pre-trained models such as BERT, RoBERTa, and Sentence-BERT, where the similarity between phrases is measured based on their semantic context. Prediction precision is calculated using cosine similarity provided by Formula 11.

$$\mathrm{Sim(PK,RK)} = \frac{\mathrm{e(PK).e(RK)}}{\|\mathrm{e(PK)}\|\|\mathrm{e(RK)}\|} \quad (11)$$

where e(·) is contextual embedding of a phrase obtained from a pretrained language model. Although this metric achieves semantic comparison, its effectiveness in keyphrase evaluation is limited because it compares the similarity of the phrase as a single unit, while lacking the ability to understand the word-for-word relationship between the prediction and the reference. To overcome this challenge, the BERTScore was proposed, which relies on contextual embedding that change depending on the word's position in the phrase. Instead to measure the cosine similarity between phrase vectors, the BERTScore aligns each token in the predicted phrase with the most semantically similar token in the reference keyphrase based on their context, thus preserving the value of each token in the keyphrase. To calculate BERTScore, precision and recall are first calculated by formulas 12 and 13.

$$\mathrm{Precision_{BERT}} = \frac{1}{|\mathrm{PK}|}\sum_{\mathrm{i=1}}^{|\mathrm{PK}|}\max_{\mathrm{j}}\cos\left(\mathrm{e(x_i)},\mathrm{e(y_j)}\right) \quad (12)$$

$$\mathrm{Recall_{BERT}} = \frac{1}{|\mathrm{RK}|}\sum_{\mathrm{j=1}}^{|\mathrm{RK}|}\max_{\mathrm{i}}\cos\left(\mathrm{e(y_j)},\mathrm{e(x_i)}\right) \quad (13)$$

where:

- e(·): The contextual embedding of a token obtained from BERT model
- cos(·): cosine similarity
- PK: Predicted keyphrase of tokens $\{x_1, \dots x_i, \dots x_{|PK|}\}$
- RK: Reference keyphrase of tokens $\{y_1, \dots y_j, \dots y_{|RK|}\}$

To calculate BERTScore, $Precision_{BERT}$ and $Recall_{BERT}$ combined using the harmonic mean via formula 14.

$$\mathrm{BERTScore} = 2\times\frac{\mathrm{Precision_{BERT}}\times\mathrm{Recall_{BERT}}}{\mathrm{Precision_{BERT}}+\mathrm{Recall_{BERT}}} \quad (14)$$

Instead of combining the words of a phrase into a single vector, BERTScore compares them word by word. Therefore, it addresses the problem of variations in phrase lengths. Also,

BERTScore evaluates semantic overlap, making it robust to paraphrasing in generative keyphrase prediction. In contrast, the precision of BERTScore is affected by the quality and comprehensiveness of the Gold Standard identified by human experts.

D. Human Reference Challenge

Manual identification keyphrases is a subjective cognitive task, where different keyphrases are selected for the same document based on the expert's background knowledge. The study [36], confirms that the manual selection of keyphrases is highly subjective. This makes relying on a single human reference limit the evaluation precision [12]. Fig. 2 illustrates the divergence between the keyphrases selected by author, Expert 1, and Expert 2.



Fig. 2. The divergence of reference keyphrases selection.

The central intersection represents the keyphrases agreed by everyone. Overlapping areas indicate partial agreement between two of them, while divergence areas represent keyphrases chosen by each independently. This representation highlights the inconsistency on gold standards to select keyphrases.

The cognitive effort expended in the manual selection, limits the number of keyphrases selected, this also, impact the evaluation process. Despite advancements in lexical and semantic evaluation metrics, the evaluation of keyphrase prediction models relies on the quality of the reference keyphrases. This leads to unreliable evaluation, especially in generative models, where they may generate semantically relevant keyphrases that are not present in the reference set.

There are several challenges to evaluate keyphrase prediction models. Current evaluation metrics remained to compare phrase-to-phrase, whether through exact or semantic matching. The keyphrases generated often represent a condensation of entire paragraphs, rather than a corresponding phrase in a human reference, which is often limited, and subjective. Therefore, we need a more comprehensive evaluation process. For this, we propose a new evaluation metric, instead of comparing predicted keyphrases with limited reference keyphrases, we compare the overall semantic impact of these phrases to generate the document summary, thus freeing the evaluation from limited human reference lists.

IV. Functional Evaluation Metric

A. Motivation for Functional Evaluation

The review of evaluation process for keyphrase prediction models in the previous section showed several challenges that limit the reliability of their results. Current evaluation metrics judge model performance solely by lexical or semantic matching between reference keyphrases and predicted keyphrases, neglecting characteristics such as topic coverage, diversity, and functional performance. Furthermore, they relie on the list of reference keyphrases, which are subjective and incomplete. Consequently, the performance of keyphrase prediction models is measured by human judgment rather than functional effectiveness, undermining the credibility of the evaluation process and penalizing models that generate functionally useful keyphrases but are not present in the reference list.

Therefore, we propose a Functional Evaluation Metric (FEF) that evaluates predicted keyphrases as a collective semantic unit, more independently of the reference list. FEF evaluates the performance of keyphrase prediction models by measuring the effectiveness of predicted keyphrases on two tracks. Abstract reconstruction track and text retrieval track, without relying on reference keyphrases. Therefore, the evaluation process is based on the functional success of the predicted keyphrases. We no longer measure whether this phrase matches the reference, but rather whether this phrase enables the system to summarize or retrieve the document.

B. Overview of FEM-KP

The Functional Evaluation Metric for KeyPhrase Prediction (FEM-KP) is a new metric designed to evaluate keyphrase prediction models based on the functional utility of the generated keyphrases, rather than measuring lexical or semantic similarity. FEM-KP measures the functional utility of generated keyphrases through two functions:

- Summary Reconstruction: This dimension measures the ability of keyphrases to serve as semantic seeds that can summarize the document and cover its content.
- Retrieval Power: This dimension measures the ability of keyphrases to identify the document within a large corpus.

The combinition of these dimensions in FEM-KP provides a uniform score that reflects the performance of the keyphrase prediction model to generate exploitable keyphrases for NLP applications such as text clustering, information retrieval, and document summarization.

C. Process of FEM-KP

To evaluate the performance of the keyphrase prediction model, the evaluation process consists of three main stages: synthesis, parallel evaluation tasks, and weight fusion. Fig. 3 illustrates how the FEM-KP score is calculated.

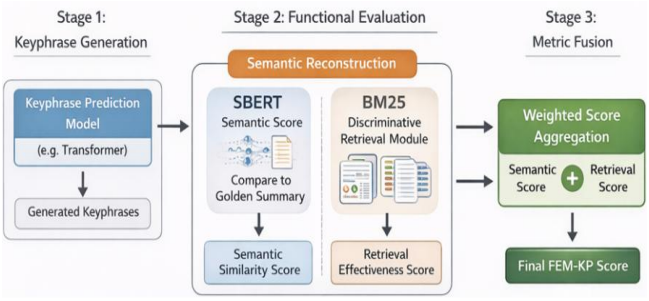


Fig. 3. Process of FEM-KP metric.

In the first stage, we use a transformer-based generator to combine and synthesize the generated keyphrases KP into a coherent semantic string SK. To achieve this, we first use a pre-trained generative model such as Text-to-Text Transfer Transformer (T5) [37] or Bidirectional and Auto-Regressive Transformers (BART) [25] for keyphrases-to-text conversion tasks.

To evaluate the ability of keyphrases to perform real tasks, FEM-KP across two parallel functional tracks. The first is the summative reconstruction score S_{summ}. In this track, we measure the semantic distance between the reconstructed summary SK and the gold summary GS using cosine similarity in the embedding space.

In contrast, the second track evaluates the ability of key phrases to recognize and retrieve the document D within a large dataset C. Its ability to handle term frequency values and normalize document length allows BM25 to calculate the retrieval score, which is an ideal metric to evaluate the retrieval utility of generated keyphrases.

The functional dimensions are combined into a single score to evaluate the overall quality of the keyphrases. The final FEM-KP score is calculated by balancing S_{summ} and S_{ret}.

D. Mathematical Formulation of FEM-KP Metric

FEM-KP framwork evaluate the quality of a keyphrase by measuring its functional utility across two NLP tasks. Let D is the source document, and $KP = \{P_1, P_2, \dots, P_n\}$ is the set of generated keyphrases from D by a keyphrase prediction model, and GS is the gold summary of D. The formulation follows a three-step pipeline:

1) Semantic reconstruction: Semantic reconstruction aims to generate a summary of the document SK based on keyphrases. T5 achieved a relative improvement in precision, but its performance is lower than BART in most summarization scenarios. For this, BART was used as a generative model in FEM-KP. To obtain the SK summary, we use the formula 15.

$$SK = arg\ \max_{s} p_\theta(s|Serialize(KP)) \quad (15)$$

where:

- SK: The generated summary.
- p_θ: The conditional probability modeled by BART
- Serialize(KP): A linearized representation of the KP set.
- s: A candidate output text sequence generated by the BART decoder.

The generate summary SK obtained from BART is represented using a Sentence-BERT [38]. The vector obtained serves as a semantic fingerprint for the generated keyphrases.

The first Track Functional Evaluation, is the ability of keyphrases to generate a document summary. In this track, we measure the semantic distance S_{summ} in vector space between the generate summary SK and the gold summary GS. To calculate semantic similarity, the SK and GS are converted into semantically vectors using SBERT model in formula 16.

$$\begin{cases} V_k = Pooling(Transformer_\theta(SK)) \\ V_g = Pooling(Transformer_\theta(GS)) \end{cases} \quad (16)$$

S_{summ} is calculated using cosine similarity in formula 17.

$$S_{summ} = \frac{\sum_{i=1}^{n} v_{k,i} \cdot v_{g,i}}{\sqrt{\sum_{i=1}^{n} v_{k,i}^2}\sqrt{\sum_{i=1}^{n} v_{g,i}^2}} \quad (17)$$

We used KP20k [39], Inspec [40], and Krapivin [41] as benchmark datasets for comparison. Their features and the reasons for their selection will be explained in the evaluation section.

2) Retrieval rank: The second track evaluates the ability of keyphrases to identify the source document D from a large set of documents C. To calculate the retrieval score, we use the BM25 probabilistic model, which improves the traditional TF-IDF model by incorporating term frequency saturation and document length normalization. Formula 18 calculates the document retrieval rank D, using the keyphrases set K.

$$Rank_K(D) = \sum_{p \in K} IDF(p).\frac{fr(p,D).(n+1)}{fr(p,D)+n.(1-\alpha+\alpha.\frac{|D|}{Avg(C)})} \quad (18)$$

where:

- $IDF(p)$: The Inverse Document Frequency of keyphrase p.
- $fr(p,D)$: The frequency of keyphrase p within document D.
- $|D|$: The length of the D.
- $Avg(C)$: The average document length across the corpus C.
- n: A hyperparameter to control the term frequency saturation. (Typically, between 1.2 and 2).
- α: A scaling factor for document length normalization. (typically, α =0.75).

It is difficult to combine the rank as score with the S_{summ} score (values between 0 and 1) in a single equation, for this normmalization is necessary. The S_{ret} score is calculated as the reciprocal rank. The retrieval score is calculate using formula 19.

$$S_{ret} = \begin{cases} \frac{1}{Rank_K(D)}, & if\ Rank_K(D) \le 100 \\ 0, & otherwise \end{cases} \quad (19)$$

The retrieval function score approaches zero if the keyphrases fail to rank the document among the top 100 results. This significantly speeds up the evaluation process, especially in datasets containing thousands of documents.

The functional dimension scores are combined into a single score to evaluate the overall performance of keyphrases. The FEM-KP score is calculated by balancing the sum of the S_{summ} and S_{ret}. This allows to adjust the score based on application priority, where a higher β Hyperparameter is used for summarization tasks, and a higher δ Hyperparameters for search optimization or indexation tasks. The final FEM-KP score was calculated using the formula 20.

$$FEF - KP = \beta . S_{summary} + \delta . S_{retrieval} \quad (20)$$

β and δ: Hyperparameters that are controlled according to the nature of the task, often $\beta + \delta = 1$.

This formula includes the effectiveness evaluation of the generated keyphrases, especially in the semantic representation of the document, and retrieval tasks.

E. Advantages and Limitations

Instead of the lexical or semantic matching used by current evaluation metrics, the FEM-KP metric adopts functional evaluation, which measures the effectiveness of keyphrases by their application in NLP tasks. The FEM-KP metric is characterized by its flexibility in adjusting the values of the hyperparameters β and γ, which allowed for the selection of the most suitable model for each task. Also, using the BM25 algorithm, the keyphrase prediction model could not achieve a high score when generating a large number of general keyphrases.

Despite these advantages, the quality of the reference summaries used may affect the reliability of the performance. Furthermore, they may not always be available. Also, using pre-trained language models and retrieval algorithms increases the computational cost of FEM-KP compared to simple lexical metrics. Also, the sensitivity of the parameters requires careful selection of β and γ values to avoid biased results

V. Experimental Setup and Results

This section details the implementation procedures and evaluation protocol for FEM-KP, including evaluation datasets and keyphrase prediction models. This section aims to ensure a comprehensive evaluation that allows for comparison of FEM-KP with current evaluation metrics.

A. Experimental Setup

1) Evaluation protocol: The evaluation protocol involves comparing the performance of key phrase prediction models using various evaluation metrics, including functional effectiveness. This comparison is achieved by calculating the BERTScore, ROUGE, and F1Measure scores for each model and comparing them to the FEM-KP scores for each model.

2) Evaluation datasets: To evaluate the FEM-KP, we rely on three reference evaluation datasets. The KP20k dataset comprises scientific abstracts from multiple disciplines. Each document has keyphrases defined by author. The size and diversity of KP20k content make it suitable for comparisons between different evaluation metrics. The second dataset, is Inspec. It smaller but more carefully curated and contains scientific abstracts. these datasets have been widely used in previous studies, making them suitable for comparing the FEM-KP with exact-match and semantic-based evaluation metrics.

TABLE II. Comparison of the Datasets Used for FEM-KP Evaluation

Dataset	Domain	# Test Documents	Avg. Text Length	Keyphrase Types
KP20k	Scientific Abstracts	20,000	180 words	Present & Absent
Inspec	Scientific Abstracts	500	135 words	Present & Absent
Krapivin	Scientific Papers	2300	8000 words	Present & Absent

To evaluate FEM-KP in long documents and complex retrieval scenarios, we exploited the Krapivin dataset, which includes scientific papers. Table II presents the features of KP20k and Inspec datasets.

3) Keyphrase prediction models: For a comprehensive comparison of the evaluation metrics, TextRank was adopted as an unsupervised, graph-based model that extracts keyphrases from the word-to-word structure within the text. CopyRNN was adopted as a supervised, recurrent neural network model that utilizes a copy mechanism to generate present and absent keyphrases in the text. Additionally, BART-based Keyphrase Generator and T5-based Keyphrase Generator were adopted, leveraging of pre-trained generative models to produce high-quality, sequential keyphrases. The diversity of models including unsupervised, supervised, and generative models, allows to evaluate FEM-KP in different scenarios. These models represent the evolution of keyphrase prediction task. Table III provides a comprehensive analysis of each model type and the rationale for its selection.

To highlight the performance of FEM-KP across different datasets for the four models and compare it to the F1-Score, ROUGE-L and BERTScore metrics. Fig. 4 summarizes the different stages of experimental evaluation.

TABLE III. A COMPREHENSIVE ANALYSIS OF KEYPHRASE PREDICTION MODELS USED IN EVALUATION

Model	Model Type	Output	Reason
TextRank [42]	Unsupervised, Graph-based	Extractive keyphrases	An unsupervised model to extract keyphrase.
CopyRNN [39]	Supervised, Neural Seq2Seq	Extractive and generative keyphrases	A neural network model to generate present and absent keyphrases.
BART-based Keyphrase Generator [25]	Transformer-based	Generative keyphrases	Abstractive model used for keyphrases prediction.
T5-based Keyphrase Generator [37]	Transformer-based (Text-to-Text)	Generative keyphrases	A metric for text transformation. Widely used in recent studies to generate keyphrases.

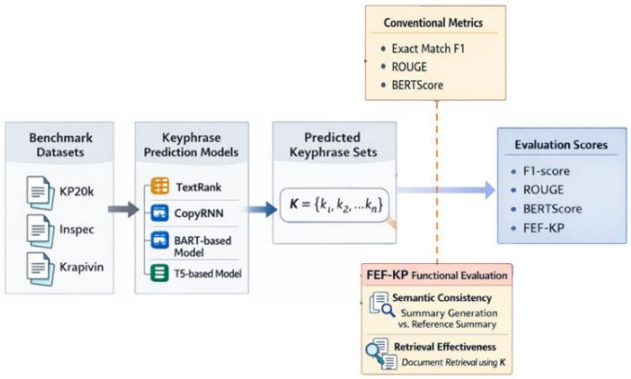


Fig. 4. The stages of experimental evaluation

We will present, analyze and discuss the evaluation results in the following part.

B. Evaluation Results

1) Sample size: To ensure the reliability of comparisons between the four models and the metrics used, samples were selected from the KP20k, Inspec, and Krapivin datasets. Table IV shows the sample size used for each dataset.

TABLE IV. THE SAMPLE SIZE USED IN EVALUATION

Dataset	Sample Size	Avg. Document Length
Inspec	500	Short (~120 words)
KP20k	2,000	Medium (~180 words)
Krapivin	400	Long (~8,000+ words)

To reduce computational costs and achieve a high level of reliability, 2,000 documents were selected from the KP20k dataset, and we used all documents from the Inspec dataset. 400 research papers were selected from the Krapivin dataset to evaluate the effectiveness retrieval of the FEM-KP in longer documents.

2) Consistency of performance ranking: The consistent ranking of models across diverse datasets is a prerequisite for any evaluation. The reliability of an evaluation metric is measured by its ability to rank model performance across diverse datasets. Despite the complexity of the dataset or the length of the document, the relative ranking of models should remain constant. In this part, we analyze the robustness of the FEM-KP metric by comparing its ranking stability for key phrase prediction models with the F1-Score, ROUGE-L, and BERTScore metrics. Our analysis aims to demonstrate that FEM-KP has a unique ability to identify the most effective model even as data complexity increases, whereas other measures are susceptible to ranking inversions and contextual noise. Table V presents the performance of the four keyphrases prediction in the Inspec dataset representing short texts.

TABLE V. THE PERFORMANCE OF THE FOUR KEYPHRASES PREDICTION IN THE INSPEC DATASET

Model	F1-Score	ROUGE-L	BERTScore	FEM-KP
TextRank	(4) 0.215	(4) 0.312	(4) 0.721	(4) 0.584
CopyRNN	(3) 0.248	(3) 0.345	(3) 0.765	(3) 0.642
T5-based	(2) 0.278	(2) 0.372	(2) 0.835	(2) 0.775
BART-based	(1) 0.285	(1) 0.380	(1) 0.842	(1) 0.791

The texts in Inspec are short summaries, for this, the potential error across all metrics is reduced. Therefore, we observe consistency across all metrics in ranking model performance. Conversely, when using the KP20K dataset, which contains longer summaries compared to Inspec, the exact Match metrics began to deviate. F1-Score favored CopyRNN over T5 and ROUGE-L favored T5 over BART, while BERTScore and FEM-KP maintained their rankings. Table VI presents the performance of the four keyphrases prediction in the KP20k dataset representing short texts.

TABLE VI. THE PERFORMANCE OF THE FOUR KEYPHRASES PREDICTION IN THE KP20K DATASET

Model	F1-Score	ROUGE-L	BERTScore	FEM-KP
TextRank	(4) 0.142	(4) 0.245	(4) 0.680	(4) 0.510
CopyRNN	**(2) 0.200**	(3) 0.290	(3) 0.730	(3) 0.615
T5-based	**(3) 0.185**	**(1) 0.325**	(2) 0.808	(2) 0.732
BART-based	(1) 0.210	**(2) 0.318**	(1) 0.815	(1) 0.748

The deviate in the ranking of exact matching metrics is due to the increasing size and diversity of the data. Additionally, these metrics are sensitive to the Longest Common String (LCS). The results also demonstrate that FEM-KP overcomes this problem through its reconstruction track. The greatest challenge arises, in Krapivin, as the metric must evaluate the performance of models in long texts. Table VII presents the performance of the four keyphrases prediction in the Krapivin dataset representing long texts.

In Krapivin, a deviate is observed in the F1-Score, where CopyRNN was favored over T5. This is due to the length of the texts, which increased the probability of random matches. Furthermore, the semantic-based BERTScore metric showed ranking instability within the Krapivin dataset. The length of

the texts also led to contextual noise resulting from the presence of irrelevant informations. Therefore, CopyRNN obtiened a high score by BERTScore because the context was perfectly matched. In contrast, FEM-KP is the only metric that maintained the rank (BART > T5 > CopyRNN > TextRank) in all datasets, proving that it is not fooled by length of the text or overlapping of phrases, but rather measures functional value.

TABLE VII. THE PERFORMANCE OF THE FOUR KEYPHRASES PREDICTION IN THE KRAPIVIN DATASET

Model	F1-Score	ROUGE-L	BERTScore	FEM-KP
TextRank	(4) 0.125	(4) 0.210	(4) 0.655	(4) 0.495
CopyRNN	**(2)0.180**	(3) 0.265	**(2) 0.785**	(3) 0.580
T5-based	**(3) 0.175**	(2) 0.302	**(3) 0.782**	(2) 0.695
BART-based	(1) 0.195	(1) 0.310	(1) 0.790	(1) 0.712

3) Discussion of results: Analysis of results from the Inspec, KP20K, and Krapivn datasets revealed differences between match evaluation and functional evaluation. While F1-Score, ROUGE-L, and BERTScore metrics showed deviate ranking with increasing document complexity and length, with extractive models like CopyRNN outperforming generative models. They often rewarded lexical redundancy and contextual matching at the expense of semantic utility. In contrast, FEM-KP demonstrated ranking consistency (BART > T5 > CopyRNN > TextRank) regardless of datasets. This stability is attributed to the dual-tracks structure of our metric. The reconstruction track rewards only the keyphrases capable to generate the document maining, while the retrieval track rewards those capable to retrieve the document. The FEM-KP result proves to be unaffected by text length and independent of the model. This provide a more reliable metric to evaluate generative keyphrase prediction models. These results confirm that the FEM-KP not only measures similarity but also functional information, making it a more reliable tool for researchers in the NLP field.

VI. CONCLUSION

In this study, we addressed the evaluation of keyphrase prediction models by highlighting the challenges posed by current evaluation metrics such as F1-Score, ROUGE-L, and BERTScore, which primarily focus on exact matching or contextual similarity but often fail to detect the functional utility of keyphrases. This leads to inconsistent model rankings, particularly in long and complex documents. Therefore, to overcome this problem, we proposed the FEM-KP metric. Our metric adopts a two-track evaluation, the utility of reconstruction (track A) and semantic consistency (track B), which has allowed to distinguish between contextually similar phrases and functionally necessary keyphrases. Our experimental results across three diverse datasets, Inspec, KP20k, and Krapivin showed that FEM-KP provides a superior and more stable evaluation metric and provide a fair assessment that rewards the functional performance of the phrases rather than literal repetition compared to current evaluation metrics. In conclusion, instead to evaluate keyphrase prediction models based to compare the predicted keyphrases with the reference keyphrases, FEM-KP allows to evaluate these models based on the functional performance of these phrases. Therefore, we believe this metric will provide a more reliable benchmark for researchers to develop highly effective NLP systems. Therefore, we will expand FEM-KP in the future to include other functions such as classification and clustering.

REFERENCES

[1] E. Papagiannopoulou and G. Tsoumakas, "A review of keyphrase extraction," WIREs Data Min & Knowl, vol. 10, no. 2, p. e1339, Mar. 2020, doi: 10.1002/widm.1339.

[2] C. Yoo and H. Lee, "Improving Abstractive Dialogue Summarization Using Keyword Extraction," Applied Sciences, vol. 13, no. 17, p. 9771, Aug. 2023, doi: 10.3390/app13179771.

[3] F. Boudin, Y. Gallina, and A. Aizawa, "Keyphrase Generation for Scientific Document Retrieval," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 1118–1126. doi: 10.18653/v1/2020.acl-main.105.

[4] J. Gabín, J. Parapar, and C. Macdonald, "Beyond Questions: Leveraging ColBERT for Keyphrase Search," 2024, arXiv. doi: 10.48550/ARXIV.2412.03193.

[5] L. T. Do, A. Bodke, P. S. Akash, and K. C.-C. Chang, "ERU-KG: Efficient Reference-aligned Unsupervised Keyphrase Generation," 2025, arXiv. doi: 10.48550/ARXIV.2505.24219.

[6] D. Wu, D. Yin, and K.-W. Chang, "KPEval: Towards Fine-Grained Semantic-Based Keyphrase Evaluation," 2023, arXiv. doi: 10.48550/ARXIV.2303.15422.

[7] H. Ding and X. Luo, "LongDocRank: graph-augmented large language models for unsupervised keyphrase extraction from long documents," J Big Data, Jan. 2026, doi: 10.1186/s40537-025-01339-8.

[8] Y. Luo, Y. Xu, J. Ye, X. Qiu, and Q. Zhang, "Keyphrase Generation with Fine-Grained Evaluation-Guided Reinforcement Learning," in Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 2021, pp. 497–507. doi: 10.18653/v1/2021.findings-emnlp.45.

[9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," 2019, arXiv. doi: 10.48550/ARXIV.1904.09675.

[10] J. Jose and B. Soundarabai, "A Survey on Machine Learning Based Keyphrase Generation in Natural Language Processing," Nepal Journal of Science and Technology, vol. 22, no. 2, pp. 66–74, Dec. 2023, doi: 10.3126/njst.v22i2.85238.

[11] L. Ajallouda, F. Z. Fagroud, A. Zellou, and E. H. Benlahmar, "A Systematic Literature Review of Keyphrases Extraction Approaches," Int. J. Interact. Mob. Technol., vol. 16, no. 16, pp. 31–58, Aug. 2022, doi: 10.3991/ijim.v16i16.33081.

[12] K. S. Hasan and V. Ng, "Automatic Keyphrase Extraction: A Survey of the State of the Art," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, 2014, pp. 1262–1273. doi: 10.3115/v1/P14-1119.

[13] L. Ajallouda, F. Z. Fagroud, A. Zellou, and E. H. Benlahmar, "Automatic keyphrases extraction: an overview of deep learning approaches," Bulletin EEI, vol. 12, no. 1, pp. 303–313, Feb. 2023, doi: 10.11591/eei.v12i1.4130.

[14] M. Song, Y. Feng, and L. Jing, "A Survey on Recent Advances in Keyphrase Extraction from Pre-trained Language Models," in Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, 2023, pp. 2153–2164. doi: 10.18653/v1/2023.findings-eacl.161.

[15] B. Xie et al., "From Statistical Methods to Deep Learning, Automatic Keyphrase Prediction: A Survey," 2023, doi: 10.48550/ARXIV.2305.02579.

[16] N. Giarelis and N. Karacapilidis, "Deep learning and embeddings-based approaches for keyphrase extraction: a literature review," Knowl Inf Syst, vol. 66, no. 11, pp. 6493–6526, Nov. 2024, doi: 10.1007/s10115-024-02164-w.

[17] M. Umair, T. Sultana, and Y.-K. Lee, "Pre-Trained Language Models for Keyphrase Prediction: A Review," 2024, doi: 10.48550/ARXIV.2409.01087.

[18] B. Kang and Y. Shin, "Empirical Study of Zero-shot Keyphrase Extraction with Large Language Models," In Proceedings of the 31st International Conference on Computational Linguistics, pp. 3670–3686, 2025.

[19] 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.

[20] A. Urlana, P. Mishra, T. Roy, and R. Mishra, "Controllable Text Summarization: Unraveling Challenges, Approaches, and Prospects - A Survey," in Findings of the Association for Computational Linguistics ACL 2024, Bangkok, Thailand, 2024, pp. 1603–1623. doi: 10.18653/v1/2024.findings-acl.93.

[21] G. Erkan and D. R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," jair, vol. 22, pp. 457–479, Dec. 2004, doi: 10.1613/jair.1523.

[22] A. M. Rush, S. Chopra, and J. Weston, "A Neural Attention Model for Abstractive Sentence Summarization," 2015, arXiv. doi: 10.48550/ARXIV.1509.00685.

[23] R. Nallapati, B. Zhou, C. Dos Santos, C. Gulcehre, and B. Xiang, "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond," in Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, 2016, pp. 280–290. doi: 10.18653/v1/K16-1028.

[24] 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.

[25] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," 2019, arXiv. doi: 10.48550/ARXIV.1910.13461.

[26] T. B. Brown et al., "Language Models are Few-Shot Learners," 2020, arXiv. doi: 10.48550/ARXIV.2005.14165.

[27] Y. Zhang, H. Jin, D. Meng, J. Wang, and J. Tan, "A comprehensive survey on automatic text summarization with exploration of LLM-based methods," Neurocomputing, vol. 663, p. 131928, Jan. 2026, doi: 10.1016/j.neucom.2025.131928.

[28] S. Mushtaq and K. Veningston, "Biomedical text summarization with large language models: methodologies, challenges, and future directions," Int J Data Sci Anal, vol. 22, no. 1, p. 29, Dec. 2026, doi: 10.1007/s41060-025-00956-z.

[29] L. Bednarczyk et al., "Scientific Evidence for Clinical Text Summarization Using Large Language Models: Scoping Review," J Med Internet Res, vol. 27, p. e68998, May 2025, doi: 10.2196/68998.

[30] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," FNT in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[31] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," 2020, arXiv. doi: 10.48550/ARXIV.2004.04906.

[32] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," 2020, arXiv. doi: 10.48550/ARXIV.2004.12832.

[33] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise Zero-Shot Dense Retrieval without Relevance Labels," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, 2023, pp. 1762–1777. doi: 10.18653/v1/2023.acl-long.99.

[34] W. X. Zhao, J. Liu, R. Ren, and J.-R. Wen, "Dense Text Retrieval based on Pretrained Language Models: A Survey," 2022, arXiv. doi: 10.48550/ARXIV.2211.14876.

[35] Y.-A. Liu, R. Zhang, J. Guo, M. de Rijke, Y. Fan, and X. Cheng, "Robust Neural Information Retrieval: An Adversarial and Out-of-distribution Perspective," 2024, arXiv. doi: 10.48550/ARXIV.2407.06992.

[36] S. N. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin, "SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles," Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 21–26, 2010.

[37] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," 2019, arXiv. doi: 10.48550/ARXIV.1910.10683.

[38] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.

[39] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, and Y. Chi, "Deep Keyphrase Generation," 2017, arXiv. doi: 10.48550/ARXIV.1704.06879.

[40] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in Proceedings of the 2003 conference on Empirical methods in natural language processing, 2003, pp. 216–223. doi: 10.3115/1119355.1119383

[41] M. Krapivin, A. Autayeu, and M. Marchese, "Large dataset for keyphrases extraction," Technical Report, University of Trento, 2010.

[42] R. Mihalcea, "Graph-based ranking algorithms for sentence extraction, applied to text summarization," in Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Barcelona, Spain, 2004, pp. 20-es. doi: 10.3115/1219044.1219064.

A Neutrosophic Machine Learning-Based Intelligent Ensemble Model for Sustainable Tea Yield Prediction Under Climatic Variability

Maitraya Dey[1], Pushpita Roy[2], Shubhendu Banerjee[3]*, Amrut Ranjan Jena[4],
Rakesh Naskar[5], Suparna Dasgupta[6], Soumyabrata Saha[7], Sudarshan Nath[8], Bikash Mondal[9]
Department of CSE, Narula Institute of Technology, Kolkata, India[1, 2, 3]
Dr. Sudhir Chandra Sur Institute of Technology and Sports Complex, Kolkata, India[4]
Department of CSE, Gargi Memorial Institute of Technology, Kolkata, India[5]
Department of Information Technology, JIS College of Engineering, Kalyani, Nadia, India[6, 7, 8]
Department of ECS, Narula Institute of Technology, Kolkata, India[9]

***Abstract*—For efficient agricultural planning, resource management, and enhancing farmer livelihoods in significant tea-producing regions, tea production prediction is essential. However, climate variability—including temperature, rainfall, humidity, and sunlight duration—has a significant impact on tea output, making precise forecasting difficult. Using meteorological data from 2015 to 2025, this study suggests a hybrid machine learning approach for predicting tea production. Initially, four models are created as separate predictors: Random Forest, XG Boost, Light GBM, and Cat Boost. Three ensemble models are shown to increase prediction accuracy: a Neutrosophic Ensemble Model, a Fuzzy Logic Weighted Ensemble, and an optimized weighted ensemble utilizing Sequential Least Squares Programming (SLSQP). According to experimental data, the optimized ensemble outperforms individual and alternative ensemble models, achieving the best performance with an R^2 value of 0.86, an RMSE value of 130.89, and an MAE value of 103.96. The suggested methodology improves the accuracy of the tea yield forecast while managing climate variability.**

***Keywords*—Random Forest; XG Boost; Light GBM; Cat Boost; SLSQP; Fuzzy; Neutrosophic**

I. Introduction

Tea is a cash crop used for plantation in the world economy and plays a very important role in the agricultural economy of many developing nations. Tea crop (Camellia sinensis) is highly sensitive to climatic parameters, such as temperature, rainfall, humidity and solar irradiance. The alteration of climatic parameters hampers the productivity of the crop, as well as the sustainability of crop cultivation in the long run. Over the past few decades, climate variability and extreme weather have impacted the sustainability of tea crop production. Research has demonstrated the effect of climatic variables on the growth patterns of tea crop production, underscoring the importance of yield forecasting for proper management of tea crop production resources [1-2].

Recently, optimized weighted ensemble models for climate-based tea yield forecasting have attracted considerable attention within agriculture analytics. That is because these models can combine many different individual prediction models to produce an improved prediction accuracy of forecasted yield. Generally, statistical models frequently used for yield forecast have historically struggled with the non-linear relationship between climate variables and tea yield. However, machine learning algorithms such as Random Forest and XG Boost have demonstrated superior performance when working with larger data sets as well as managing the non-linear relationship of climate variables to tea yield than traditional regression-style forecasting models [3-4].

The climate greatly affects tea cultivation performance. This includes things like how hot or cold the weather is, how much rain falls in a specific region, and when the various growing periods occur within the season. There has been extensive research done in regard to how climatic conditions impact the amount of tea (i.e., tea yield) produced [5-6]. Climate change is also seen to be a key driving force that influences tea yields across many countries, including India, Kenya, Sri Lanka and China, as well as across other regions [7-8]. For instance, climatic conditions in regions where tea is grown, like Assam, play an important role in determining how easy or difficult it will be to produce large amounts of quality tea (i.e., the trend of tea yields), therefore requiring the development of sophisticated modelling techniques to assess what the potential yields will be for future growing seasons within those regions [2]. Likewise, variations in the way that rainfall patterns and temperature affect how much tea will be produced in each area where tea is produced are also dramatic.

Fuzzy set theory and Neutrosophic logic are two soft computing methods designed to help deal with the uncertain and imprecise nature of agricultural data. These methods can be used to consider incomplete, inconsistent, and uncertain data that frequently occur in agricultural and climatic datasets [9-10], thus enabling the use of these methods in predictive models, which not only strengthen the robustness of these models, but also improve their reliability when used in an applied setting. The advent of large-scale agricultural and climatic datasets has facilitated the development of data-driven forecasting models. There are significant historical datasets in the International Statistical Database that may serve as useful tools for the analysis of crop yield forecasting and for the development of

*Corresponding author

climatic forecasting models based primarily on sophisticated machine learning techniques [11-12].

To explore future challenges and provide opportunities in terms of establishing an optimized ensemble method using weights for climate-based tea yield prediction models, this study intends to find solutions. The authors have developed a prototype combination of various machine learning models while assigning optimal weights in order to create the best-performing model of its type. The purpose of this model is to examine how climatic influences combined with ensemble approaches can produce the most accurate estimates when predicting tea yields. It is hoped that such an approach will assist agriculture policymakers, tea planters and researchers with the decision-making process related to the management of tea as it relates to climatic impacts and sustainability.

II. Literature Review

The production of tea can be affected by several climatic and environmental components, which various modeling techniques have examined. Raj et al. [13] analyzed the long-term variation of tea production in some India tea growing areas using data from 1981 to 2015. Different statistical and time series techniques (like multiple regression, SARIMAX and artificial neural networks [ANNs]) were utilized and the authors concluded that the ANN and vector autoregressive model would be best to show Nonlinear interactions between climate input and tea yield, as well as identifying temperature variation as the primary driver for tea yield.

Batool et al. [14] developed a hybrid method for estimating tea yield through a combination of the FAO AquaCrop simulation model and machine learning models. Using weather, soil, crop, and management data from Pakistan, they demonstrated that predictive errors with machine learning (especially with the XG Boost algorithm) were lower than those associated with the simulation method, thus showing great promise for the prediction of tea yield using data-driven models.

In a study by Xiang et al. [15], ecological variables were controlled to determine the effect of temperature, humidity, light intensity, and substrate moisture on physiological function in tea plants. The study results indicate that some combinations of ecological variables exert a significant synergistic influence over shoot development, yield, and total catechin content of tea foliage, supporting a correlation between tea yield and quality. Raghav et al. [16] employed Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) networks for forecasting tea area, production and yield based on historical data from 1918 to 2023. Results of the study showed that the CNN model had lower error values than the LSTM model, which indicates that the CNN model has superior predictive capability relative to the LSTM model. This study illustrates that deep learning models are capable of accurately forecasting tea yield. Premkumar et al. [17] examined climate variability as it relates to tea production in Tamil Nadu using an ARDL error correction model. The authors concluded that lower temperatures and higher relative humidity promote tea production; however, rainfall negatively affects tea production.

The tea yield is known to be highly sensitive to both climate and environment, and recent research indicates that the traditional statistical models often do not adequately capture the complex or nonlinear relationship between climate and tea yield. While some advanced methods have demonstrated better success, such as machine learning, hybrid simulation/machine learning models, and controlled ecological experiments, there remain limitations concerning the applicability of these methods both regionally and operationally. In addition, temperature, rainfall and humidity - all identified as climatic variables relevant to tea cultivation - vary widely throughout the world; therefore, area-specific studies must be conducted. In light of these issues, this study will focus on developing an accurate and robust modelling method for predicting tea yields using climatic variables and advanced machine learning techniques.

III. Methodology

Fig. 1 represents an illustration of the proposed methodology's overview. In this example, we show how a machine learning-based approach provides predictions on tea production within India. First, a comprehensive dataset that is publicly available and encompasses a ten-year period and five major Indian states that are predominantly responsible for producing tea has been compiled. The study incorporates datasets collected across diverse temporal and environmental conditions, which inherently capture varying climatic scenarios and thus indirectly validate the model's robustness to climatic variability. The dataset will be split into model evaluation and model training datasets for two separate analyses. Model evaluation will be done by utilizing individual models against each data split during the model-building phase. The models that will be employed to develop the prediction model will consist of Random Forest (RF), XG Boost (XGB), Light GBM (LGBM), and Cat Boost models, which will be trained respectively on the newly created evaluation datasets. Further, an ensemble will be developed based on the combination of all four models trained to yield a better prediction for tea yield. The average weight applied to each of the models will be statistically determined based on the degree to which its prediction was accurate compared to the actual tea production on an annual basis, throughout the combined evaluation. The weight assigned to each model will then serve as the basis for developing the predictions for the current model based on this methodology. Algorithm 1 identifies the steps included in the proposed methodology.

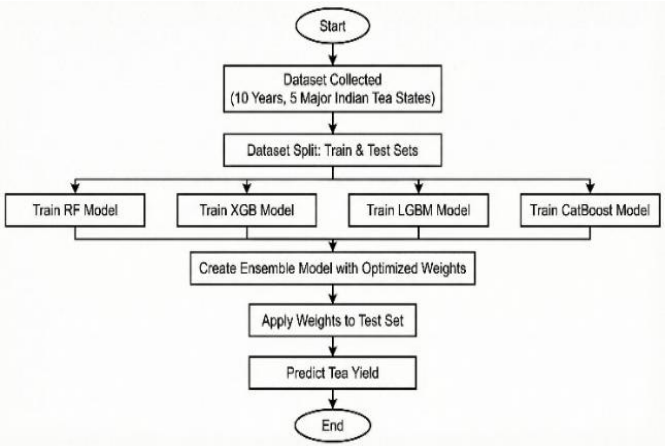


Fig. 1. Flow chart of proposed methodology.

Algorithm 1: SLSQP-based Optimized Weighted Ensemble Model

Require: Dataset D_{total}, Minimum weight threshold δ

Ensure: Final ensemble prediction $\hat{Y}_{final}$

Import training and test data

Split dataset into feature matrix X and target vector y

Perform one-hot encoding on categorical variables (State, District)

Split data into D_{train} and D_{val}

Train regression models: RF, XGB, LGBM, CAT on D_{train}

For each model $m_k \in \{RF, XGB, LGBM, CAT\}$ do

$$\hat{y}_{val}^{(k)} = m_k(X_{val})$$

End for

Form prediction matrix $P \in \mathbb{R}^{N_{val} \times 4}$:

$$P = \left[\hat{y}_{val}^{(1)}, \hat{y}_{val}^{(2)}, \hat{y}_{val}^{(3)}, \hat{y}_{val}^{(4)}\right]$$

For each model m_k do

$$RMSE_k = \sqrt{\frac{1}{N_{val}} \sum_{i=1}^{N_{val}} \left(y_i - \hat{y}_i^{(k)}\right)^2}$$

End for

Initialize weights using inverse RMSE:

$$w_k^{(0)} = \frac{1/(RMSE_k)}{\sum_{j=1}^{4} 1/(RMSE_j)}$$

Ensemble prediction:

$$\hat{y}_i = \sum_{k=1}^{4} w_k P_{i,k}$$

Objective function:

$$\min_w J(w) = \sqrt{\frac{1}{N_{val}} \sum_{i=1}^{N_{val}} (y_i - \hat{y}_i)^2}$$

Subject to equality constraint:

$$\sum_{k=1}^{4} w_k = 1, \qquad w_k \geq \delta$$

Apply Sequential Least Squares Programming (SLSQP)

Iteratively update weight vector w using quadratic approximation while satisfying constraints

Obtain optimal weight vector w^*

For each model m_k do

Retrain m_k on full dataset D_{total}

$$\hat{y}_{test}^{(k)} = m_k(X_{test})$$

End for

$$\hat{Y}_{final} = \sum_{k=1}^{4} w_k^* \hat{y}_{test}^{(k)}$$

Return $\hat{Y}_{final}$

Algorithm 1 describes the overall structure of the traditional optimized weighted ensemble paradigm, wherein multiple machine learning models are trained and the optimal weights are determined via constrained optimization techniques based on the validation data set. Each individual model is trained separately on the same training data and produces a corresponding prediction $\hat{y}_i$ for a given input. Initial weights are derived using the inverse Root Mean Square Error (RMSE) of each base model:

$$w_i^{(0)} = \frac{1/(RMSE_i)}{\sum_{j=1}^{M} 1/(RMSE_j)} \quad (1)$$

The ensemble weights are optimized by minimizing the RMSE of the ensemble on the validation set using the Sequential Least Squares Programming (SLSQP) optimization algorithm. The objective function is defined as:

$$\min_w \sqrt{\frac{1}{n} \sum_{k=1}^{n} \left(y_k - \sum_{i=1}^{M} w_i \hat{y}_{i,k}\right)^2} \quad (2)$$

The final ensemble prediction is computed as a weighted linear combination of the base model predictions, expressed as:

$$\hat{y} = \sum_{i=1}^{M} w_i \hat{y}_i \quad (3)$$

where, M denotes the number of base models and w_i represents the weight assigned to the i^{th} model. To ensure a valid convex combination, the weights are constrained to satisfy:

$$\sum_{i=1}^{M} w_i = 1, w_i \geq 0 \quad (4)$$

Each of the base learners is trained separately on the same training dataset and generates predicted values for the tea yield. The prediction error of each learner is calculated on the validation dataset using the Root Mean Square Error (RMSE) metric. The RMSE values are used as crisp inputs to the fuzzy logic system.

Fuzzification Using Gaussian Membership Function: Each RMSE value is converted to a fuzzy membership degree, which represents the level of trust in the respective model. A Gaussian membership function is used:

$$\mu_i = exp\left(-\frac{RMSE_i^2}{2\sigma^2}\right) \quad (5)$$

where, σ is the adaptive spread parameter, which is calculated as the mean RMSE of all base models:

$$\sigma = \frac{1}{4} \sum_{i=1}^{4} RMSE_i \quad (6)$$

The membership degree increases with lower RMSE values, representing higher confidence in the model.

Defuzzification: The fuzzy membership degrees are normalized to get the final ensemble weights:

$$w_i = \frac{\mu_i}{\sum_{j=1}^{4} \mu_j} \quad (7)$$

The final fuzzy ensemble prediction for the k[th] sample is calculated as a weighted sum of the individual predictions:

$$\hat{y}_k = \sum_{i=1}^{4} w_i \hat{y}_{i,k} \quad (8)$$

These values confirm that the fuzzy logic-based weighting approach successfully combines the predictions of the individual base models, providing accurate and robust predictions of tea yield even in uncertain conditions.

Extending this traditional methodology further, Algorithm 2 introduces the concept of fuzzy logic to determine the adaptive weights of the machine learning models based on the performance on the validation dataset.

Algorithm 2: Proposed Fuzzy Logic Weighted Ensemble (FWE) Algorithm

Require: Validation dataset $D_{val} = \{(x_i, y_i)\}_{i=1}^{n}$, Base models $M = \{m_1, m_2, \dots, m_k\}$

Ensure: Final ensemble prediction Y_{ens}

For each model $m_j \in M$ do

$\hat{y}_{i,j} = m_j(x_i)$ for all $(x_i, y_i) \in D_{val}$

Compute RMSE:

$$RMSE_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{i,j}\right)^2}$$

End for

Compute adaptive spread parameter:

$$\sigma = \frac{1}{k}\sum_{j=1}^{k} RMSE_j$$

For each model $m_j \in M$ do

Compute Gaussian membership degree:

$$\mu_j = exp\left(-\frac{RMSE_j^2}{2\sigma^2}\right)$$

End for

For each model $m_j \in M$ do

$$W_j = \frac{\mu_j}{\sum_{l=1}^{k}\mu_l}$$

End for

For a new input x_{new} do

$$Y_{ens}(x_{new}) = \sum_{j=1}^{k} W_j \cdot m_j(x_{new})$$

End for

Return Y_{ens}

Algorithm 2 shows a Fuzzy Logic Weighted Ensemble (FWE) method in which the weights of models are computed using Gaussian membership functions, which are functions of the corresponding RMSE values. In this approach, the contribution of each base learner in the ensemble output is adaptively controlled in proportion to the accuracy of the base learner, so that models with higher accuracy contribute more to the ensemble output. However, fuzzy logic essentially represents membership functions without considering any uncertainty or indeterminacy in the system output. Therefore, to improve the robustness of the fuzzy-based approach, Algorithm 3 proposes a Neutrosophic Weighted Ensemble (NWE) method, in which a fuzzy-based approach is generalized to consider all three components of a Neutrosophic set, namely, truth (measured in terms of R^2 values), indeterminacy, and falseness (measured in terms of normalized RMSE values), to obtain a more robust and reliable ensemble output, considering all the uncertainty in the system output, to obtain a more informative ensemble output compared to the fuzzy-based approach.

The Neutrosophic approach defines uncertainty using three parts: Truth (T), Indeterminacy (I), and Falsity (F). The values of T, I, and F are calculated for each base model using the validation performance metrics.

Truth Membership (T): Truth reflects the predictive reliability of a model and is defined using the coefficient of determination.

$$T_i = R_i^2 \tag{9}$$

Indeterminacy Membership (I): Indeterminacy captures model uncertainty.

$$I_i = 1 - R_i^2 \tag{10}$$

Falsity Membership (F): Falsity represents prediction error and is derived from the normalized RMSE.

$$F_i = \frac{RMSE_i - min\,(RMSE)}{max(RMSE) - min\,(RMSE)} \tag{11}$$

A single Neutrosophic score is computed through deneutrosophication to balance accuracy, uncertainty, and error.

$$S_i = \frac{2 + T_i - I_i - F_i}{3} \tag{12}$$

The final ensemble weight for each model is obtained by normalizing these scores.

$$w_i = \frac{S_i}{\sum_{j=1}^{4} S_j} \tag{13}$$

The Neutrosophic ensemble prediction:

$$\hat{y}_k = \sum_{i=1}^{4} w_i \hat{y}_{i,k} \tag{14}$$

The performance metrics show that the combination of Neutrosophic logic helps to strike a balance between the accuracy of the model and the uncertainty of the environment, resulting in a well-performing ensemble model for tea yield prediction. These three algorithms, in total, form a progressive and robust ensemble learning methodology. Algorithm 1 provides a solid foundation by optimally combining multiple machine learning models through constrained optimization, thus achieving balanced weight allocation mathematically. Algorithm 2 extends this methodology by incorporating fuzzy logic, thus achieving adaptive and performance-sensitive weighting according to validation errors, thus increasing the flexibility of the methodology. Algorithm 3 extends this methodology by incorporating Neutrosophic logic, thus achieving more efficient management of uncertainty and variability in the performance of models, thus achieving a more reliable, intelligent, and robust ensemble prediction system, moving from a deterministic approach to uncertainty management to intelligent weighting.

Algorithm 3: Proposed Neutrosophic Weighted Ensemble (NWE) Algorithm

Require: Validation dataset $D_{val} = \{(x_i, y_i)\}_{i=1}^{n}$, Base models $M = \{m_1, m_2, \dots, m_k\}$

Ensure: Final ensemble prediction Y_{ens}

For each model $m_j \in M$ do

Compute validation metrics: R_j^2 and $RMSE_j$ on D_{val}

End for

$E_{min} = min(RMSE_j)$
$E_{max} = max(RMSE_j)$
For each model $m_j \in M$ do
  Truth membership: $T_j = R_j^2$
  Indeterminacy membership: $I_j = 1 - R_j^2$
  Falsity membership:

$$F_j = \frac{RMSE_j - E_{min}}{E_{max} - E_{min}}$$

End for
For each model $m_j \in M$ do
  Compute neutrosophic score:

$$S_j = \frac{2 + T_j - I_j - F_j}{3}$$

End for
Normalize scores to obtain weights:

$$W_j = \frac{S_j}{\sum_{l=1}^{k} S_l}$$

For a new input instance x_{new} do

$$Y_{ens}(x_{new}) = \sum_{j=1}^{k} W_j \cdot m_j(x_{new})$$

End for
Return Y_{ens}

IV. Results

A. Dataset Description

The empirical data are in numerical form, representing observations of tea production (kg/ha) and other agro-meteorological factors of major tea-yielding states like Assam, Himachal Pradesh, Kerala, Tamil Nadu, West Bengal, recorded at the district level during the period 2015 to 2025. The dataset is taken from [12]. Table I shows the description of the dataset that includes temporal factors and other climatic variables such as temperature, rainfall, relative humidity, and sunshine hours. The data is presented in arithmetic mean values, along with the minimum and maximum ranges, thus reflecting the variability of the environmental factors that affected tea production during the period. The data is given as a bar chart in Fig. 2.

TABLE I. Dataset Description

Variables	Mean (min-max)
Tea Yield (kg/ha)	1750.27 (1021.0–2923.4)
Maximum Temperature (°C)	29.09 (19.2–37.8)
Minimum Temperature (°C)	19.25 (2.6–26.5)
Rainfall (mm)	119.48 (0.0–485.6)
Relative Humidity – Morning (%)	85.24 (65.0–100.0)
Relative Humidity – Afternoon (%)	66.91 (34.7–96.7)
Sunshine Hours (hours)	5.60 (0.2–11.5)

The input of these regression algorithms is meteorological factors, and the target value is the observed tea yield. Data is split into a ratio of 80% and 20%. We have utilized 80% of the data for training the models, and the rest of 20% is used for testing.

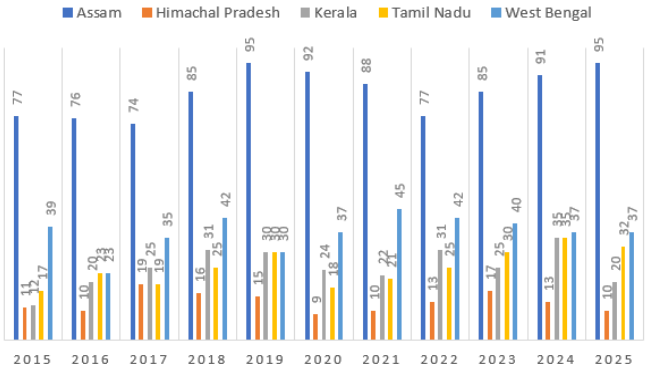


Fig. 2. Year-wise data collection

B. Performance of the Machine Learning Techniques for Tea Yield Prediction

Here we use four initial machine learning models, Random Forest, XG Boost, LGBM, Cat Boost for implementation, and also use a weighted ensemble model for better results. The performance metrics are as follows:

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \quad (15)$$

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum(y-\hat{y})^2} \quad (16)$$

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum|y-\hat{y}| \quad (17)$$

In the above equations, y is the observed yield and $\hat{y}$ is the predicted yield, and n is the total number of samples. The performance of all the machine learning models and the ensemble model is given in Table II and Fig. 3.

TABLE II. Performance of Machine Learning Techniques

Model	R^2	RMSE	MAE
Random Forest	0.84	139.69	111.07
XG Boost	0.83	141.92	112.49
Light GBM	0.83	144.26	116.09
Cat Boost	0.85	131.94	104.03
SLSQP-based Ensemble Model	0.86	130.89	103.96

Table II shows the comparative assessment of the models, with each model's performance compared and evaluated using R^2, RMSE, and MAE. From the table, it is clear that the Cat Boost model outperforms the Random Forest, XG Boost, and Light GBM models, as it has the highest R^2 and lower error values. However, the Ensemble Model, which is based on Sequential Least Squares Programming, has the highest R^2 value of 0.86 and the least error values of 130.89 and 103.96,

respectively. This shows that the ensemble model is more reliable compared to the other models. The ensemble model is more effective compared to the other models. Optimized weights are: Random Forest: 0.26, XG Boost: 0.01, Light GBM: 0.02, Cat Boost: 0.69.

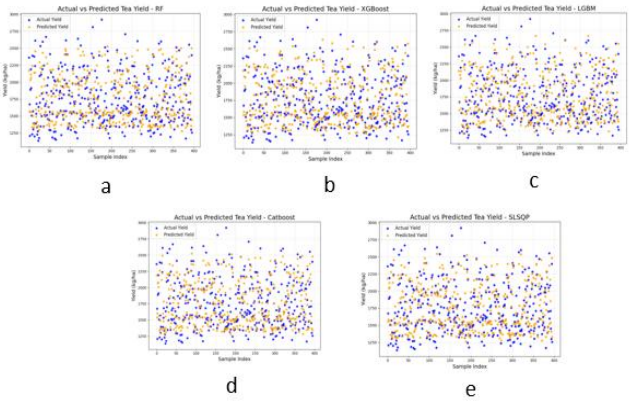


Fig. 3. Graphical representation of machine learning models: a) Random Forest, b) XG Boost, c) Light GBM, d) Cat Boost, e) Weighted ensemble model.

C. Performance for Fuzzy Logic for Tea Yield Prediction

To overcome the uncertainty associated with the prediction of tea yield and to apply dynamic weights to the learners, an ensemble framework using fuzzy logic is developed. The proposed ensemble framework combines four tree-based regression learners: Random Forest (RF), XG Boost (XGB), Light GBM (LGBM), and Cat Boost (CAT). Each of the base learners is trained separately on the same training dataset and generates predicted values for the tea yield. The prediction error of each learner is calculated on the validation dataset using the Root Mean Square Error (RMSE) metric. The RMSE values are used as crisp inputs to the fuzzy logic system. Table III and Fig. 4 show the performance of the Fuzzy-Based Ensemble Model.

TABLE III. PERFORMANCE OF FUZZY ENSEMBLE MODEL TECHNIQUES

Model	R^2	RMSE	MAE
Fuzzy-based Ensemble Model	0.84	136.32	108.77

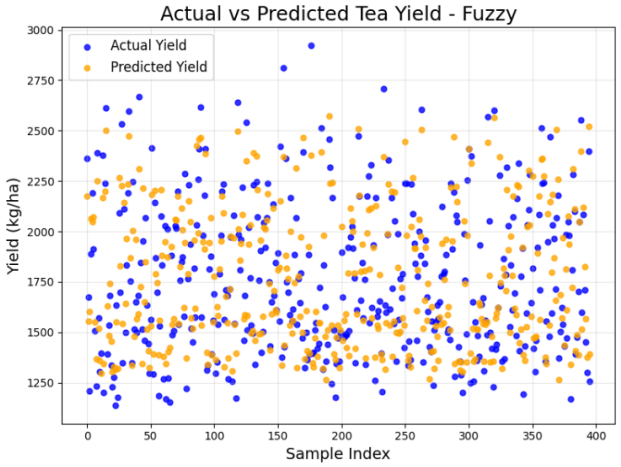


Fig. 4. Graphical representation of fuzzy logic-based tea yield prediction.

The model provides an R^2 value of 0.84979, which shows satisfactory accuracy for the model's predictions. The RMSE value of 136.3297 and the MAE value of 108.7715 show that the errors associated with the predictions are relatively low. Overall, the fuzzy ensemble model shows good performance and improves the accuracy of predictions by optimally assigning the weights according to the performance of the models. Optimized weights are: Random Forest: 0.2512, XG Boost: 0.2472, Light GBM: 0.2431, Cat Boost: 0.2585.

D. Performance for Neutrosophic Logic for Tea Yield Prediction

To enhance the accuracy of tea yield prediction in an uncertain environment, a Neutrosophic ensemble learning approach is proposed by combining four tree-based regression models: Random Forest (RF), XG Boost (XGB), Light GBM (LGBM), and Cat Boost (CAT). Each base model is trained separately on the same training data and produces predicted values of the tea yield, represented as $\hat{y}_i$.

The Neutrosophic approach defines uncertainty using three parts: Truth (T), Indeterminacy (I), and Falsity (F). The values of T, I, and F are calculated for each base model using the validation performance metrics, and the metrics are shown in Table IV.

TABLE IV. NEUTROSOPHIC METRICS

Model	T	I	F
RF	0.8423	0.1577	0.4725
XG Boost	0.8372	0.1628	0.7301
LGBM	0.8318	0.1682	1.0000
Cat Boost	0.8514	0.1486	0.0000

A single Neutrosophic score is computed through deneutrosophication to balance accuracy, uncertainty, and error.

$$S_i = \frac{2+T_i-I_i-F_i}{3} \quad (18)$$

The final ensemble weight for each model is obtained by normalizing these scores.

$$w_i = \frac{S_i}{\Sigma_{j=1}^{4} S_j} \quad (19)$$

The performance of the Neutrosophic fuzzy model is shown in Table V and Fig. 5.

TABLE V. PERFORMANCE OF NEUTROSOPHIC ENSEMBLE MODEL TECHNIQUES

Model	R^2	RMSE	MAE
Neutrosophic Ensemble Model	0.85	135.94	108.32

Table V shows the overall performance of the Neutrosophic Ensemble Model. From the table, it can be concluded that the model has an R^2 value of 0.85064, which shows high predictive accuracy. At the same time, the model has an RMSE value of 135.9409 and an MAE value of 108.3259, which shows that the errors obtained from the model are quite low. Therefore, the Neutrosophic ensemble model has shown good and reliable performance while dealing with the uncertainty that occurs

while evaluating the model and computing the weights. Optimized weights are: Random Forest: 0.2596, XG Boost: 0.2281, Light GBM: 0.1952, Cat Boost: 0.3171.

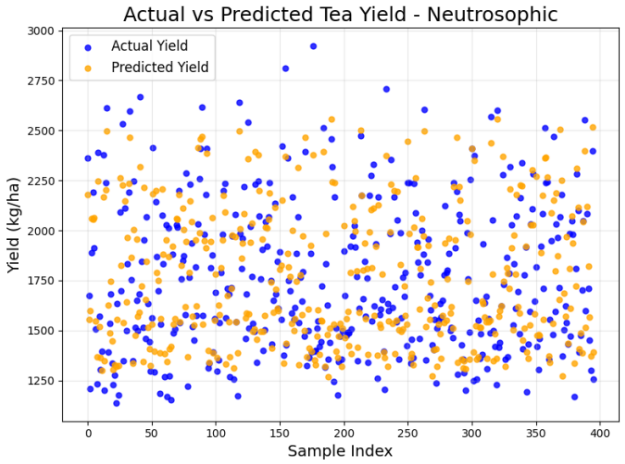


Fig. 5. Graphical representation of Neutrosophic-based tea yield prediction.

Performance of our proposed model (Sequential Least Squares Programming (SLSQP) algorithm) is improved and the results are found to be the best, with the R^2 value being high (0.8629) and the error values being low. The Fuzzy Ensemble Model and the Neutrosophic Ensemble Model are found to have high performance. The performance of the Neutrosophic Ensemble model is slightly better than the Fuzzy Ensemble model. This is evident by the R^2 value being high and the error values being low. Overall, the results show that the performance of the Ensemble model is high compared to the individual algorithms. The Ensemble model is found to have the most precise and reliable performance. The comparative analysis of our proposed ensemble model with the methods in the earlier literature is given in Table VI.

TABLE VI. COMPARISON STUDY WITH EARLIER LITERATURE

References	R^2	RMSE	MAE
Raj et al. [13]	0.612	--	--
Batool et al. [14]	--	0.154	0.123
Xiang et al. [15]	0.70	--	--
Raghav et al. [16]	--	1259.1	1086.4
Premkumar et al. [17]	0.61	--	--
Proposed Method	**0.86**	**130.89**	**103.96**

V. CONCLUSION

This study proposes various machine learning and ensemble techniques for predicting tea yield using climatic data. The proposed methodology, the optimized weighted ensemble method based on Sequential Least Squares Programming (SLSQP) is found to provide the highest prediction accuracy compared to other machine learning models. However, the models. Fuzzy and Neutrosophic ensemble models are proposed to improve the reliability of prediction techniques based on uncertainty handling capabilities. The Neutrosophic model is found to perform slightly better compared to the fuzzy model for handling uncertainty conditions.

Hence, our proposed methodology provides better prediction accuracy than Fuzzy and Neutrosophic-based models. This study can be extended by incorporating climatic data from various tea-producing states of world-wide to improve the proposed prediction techniques using machine learning models. Deep learning techniques can also be explored to improve the prediction accuracy of the proposed prediction techniques.

ACKNOWLEDGMENT

The authors sincerely acknowledge the support of JIS Group for providing financial assistance to carry out this research work. The authors are also grateful for the institutional support, infrastructure, and research environment that facilitated the successful completion of this study. Special thanks are extended to all individuals and collaborators who contributed their time and expertise toward this work.

REFERENCES

[1] S. Ahmed, J. R. Stepp, C. Orians, T. Griffin, C. Matyas, A. Robbat, S. Cash, E. Kennelly, "Effects of extreme climate events on tea (Camellia sinensis) functional quality validate indigenous farmer knowledge and sensory preferences in Tropical China", PLoS One, Vol. 9, No. 10, 2014, Article e109126.

[2] J. M. A. Duncan, S. D. Saikia, N. Gupta, E. M. Biggs, "Observing climate impacts on tea yield in Assam, India", Applied Geography, Vol. 77, 2016, pp. 64–71.

[3] L. Breiman, "Random forests", Machine Learning, Vol. 45, No. 1, 2001, pp. 5–32.

[4] T. Chen, C. Guestrin, "XGBoost: A scalable tree boosting system", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[5] M. K. V. Carr, "The climatic requirements of the tea plant: A review", Experimental Agriculture, Vol. 8, 1972, pp. 1–14.

[6] M. K. V. Carr, W. Stephens, "Climate, weather and the yield of tea", in Tea, Dordrecht: Springer Netherlands, 1992, pp. 87–135.

[7] M. Wijeratne, A. Anandacoomaraswamy, M. Amarathunga, J. Ratnasiri, B. Basnayake, N. Kalra, "Assessment of impact of climate change on productivity of tea (Camellia sinensis L.) plantations in Sri Lanka", Journal of the National Science Foundation of Sri Lanka, Vol. 35, 2007, pp. 119–126.

[8] J. Sitienei, S. G. Juma, E. Opere, "On the use of regression models to predict tea crop yield responses to climate change: A case of Nandi East, sub-county of Nandi county, Kenya", Climate, Vol. 5, 2017, Article 54.

[9] L. A. Zadeh, "Fuzzy sets", Information and Control, Vol. 8, No. 3, 1965, pp. 338–353.

[10] F. Smarandache, "Neutrosophic logic and its applications", American Research Press, 2010.

[11] FAO, "FAOSTAT statistical database", Food and Agriculture Organization of the United Nations, 2022.

[12] Biswas, "All Over India TEA YIELD dataset", Kaggle Dataset, Available: https://www.kaggle.com/datasets/arpanbiswas001/all-over-india-tea-yield

[13] E. E. Raj, K. V. Ramesh, R. Rajkumar, "Modelling the impact of agrometeorological variables on regional tea yield variability in South Indian tea-growing regions: 1981–2015", Cogent Food & Agriculture, Vol. 5, No. 1, 2019, Article 1581457.

[14] D. Batool, M. Shahbaz, H. S. Asif, K. Shaukat, T. M. Alam, I. A. Hameed, Z. Ramzan, A. Waheed, H. Aljuaid, S. Luo, "A hybrid approach to tea crop yield prediction using simulation models and machine learning", Plants, Vol. 11, No. 15, 2022, Article 1925.

[15] P. Xiang, Q. Zhu, M. Tukhvatshin, B. Cheng, M. Tan, J. Liu, J. Lin, "Response of shoot growth to ecological factors highlights a synergistic relationship between yield and catechin accumulation in tea plant (Camellia sinensis L.)", Horticulturae, Vol. 11, No. 6, 2025, Article 624.

[16] Y. S. Raghav, O. A. Alqasem, N. Mishra, B. Kumar, G. Chandel, P. Mishra, K. M. Alakdari, "Forecast using LSTM-CNN model (area, production and yield rate) of tea in India", Journal of Animal & Plant Sciences, Vol. 35, No. 3, 2025, pp. 769–779.

[17] Premkumar, R. Kishan, D. Kalaiarasi, "Repercussions of climatic variabilities on tea production in Nilgiris district of Tamil Nadu, India", Journal of Agrometeorology, Vol. 27, No. 2, 2025, pp. 173–176.

Incorporating Generative AI in Foreign Language Teaching Preparation: An Empirical Investigation Into the Effects of AI-Assisted Lesson Planning

Xinxin Guo[1], Hao Chen[2]*
Department of Public Course Teaching, Nanyang Vocational College of Agriculture, Nanyang 473000, China [1]
Artificial Intelligence Engineering College, Nanyang Vocational College of Agriculture, Nanyang 473000, China [2]

***Abstract*—Generative artificial intelligence, particularly large language models such as ChatGPT, has emerged as a promising educational technology tool with extensive application potential in teaching preparation. The research objectives of this study are: 1) to examine how generative AI affects foreign language teachers' teaching preparation experiences in terms of curriculum design, instructional resource development, teaching activity planning, work motivation, and efficiency; and 2) to identify the strengths, challenges, and teacher-recommended improvements associated with generative AI use in foreign language teaching preparation. This qualitative case study explored how generative AI assists foreign language teachers in their teaching preparation work, addressing a research gap regarding the benefits and challenges of this innovative approach. The study involved 13 foreign language teachers at a comprehensive university in central China who participated in a four-week training program. Qualitative data from semi-structured interviews were analyzed using thematic analysis. The findings indicate that generative AI positively affects teachers' preparation experiences in curriculum design, instructional resource development, and teaching activity planning, while enhancing motivation and efficiency. In degree-of-significance terms, the positive impact was large for curriculum design (reported by 10/13 teachers, approximately 77%), large for instructional resource development (10/13, approximately 77%), moderate-to-large for teaching activity planning (8/13, approximately 62%), large for work motivation (9/13, approximately 69%), and large for work efficiency (10/13, approximately 77%); by contrast, the impact on assessment design and teaching research was small (2/13 and 1/13, respectively). These insights contribute to understanding the utility and constraints of employing generative AI in foreign language teaching preparation and inform the development of effective teaching support strategies and training programs.**

***Keywords*—Artificial intelligence; ChatGPT; generative AI; educational technology; teacher professional development**

I. Introduction

Since its release in November 2022, ChatGPT, an advanced artificial intelligence language model developed by OpenAI, has garnered widespread attention in the educational field [1]. Built upon the Generative Pre-trained Transformer-3 (GPT-3) architecture with 175 billion parameters [2], ChatGPT generates human-like text responses [3] and demonstrates extensive natural language processing capabilities including query resolution, narrative generation, logical assessment, code debugging, and machine translation [4]. As a general-purpose conversational agent, ChatGPT has potential applications across education, healthcare, software engineering, content creation, and business operations [5].

Existing literature reveals growing interest in generative AI's influence on educational preparation and teacher professional development. Integration of these technologies into teaching preparation demonstrates varied strategies and perceptions, showing benefits in processing pedagogical tasks [6] alongside challenges in addressing deeper instructional design issues [7]. Teacher perceptions present a nuanced picture, with research highlighting AI's potential to enhance pedagogical experience through comprehensive feedback [8], though faculty opinions on its effectiveness vary [9]. Sociocultural factors significantly influence educators' attitudes toward AI [10], emphasizing the importance of contextual considerations in AI adoption.

Generative AI integration into curriculum design and teaching methods suggests innovation potential. Natural language processing applications [11] and teacher-AI collaboration [12] indicate forward-thinking approaches to curriculum development, underscoring the need for supportive environments to maximize AI benefits in teaching preparation. AI shows promise for enhancing instructional resource development and activity design [13], with ChatGPT and prompt engineering proving effective in automatic question generation [14]. However, limitations in identifying deep pedagogical errors [7] necessitate balanced approaches that leverage AI capabilities while recognizing human expertise. Research indicates educators prioritize practical applicability of AI-generated feedback [15], and AI-based systems show utility in mapping learning outcomes and quality assurance [16].

Challenges and ethical considerations are significant to the discourse on AI in education. Concerns regarding academic integrity, the pedagogical implications of AI-generated texts [17], and the difficulties faced by educators in identifying AI-generated content highlight the complexities of incorporating AI into teaching preparation frameworks. Furthermore, the potential impact of AI on professional role devaluation among educators necessitates a strategic approach for its adoption [18]. Finally, the potential and limitations of generative tools such as ChatGPT in enhancing teaching capabilities are evident. While such technologies offer valuable applications and can improve instructional quality [19], ethical concerns and the balance

*Corresponding author.

between innovation and responsible use remain critical considerations [20]. The literature thus delineates a landscape where AI's contributions to teaching preparation are significant, yet circumscribed by challenges that require careful navigation to ensure ethical, effective, and inclusive educational practices.

Despite substantial research on AI applications in language learning, empirical studies examining how generative AI supports foreign language teachers' preparation work remain scarce. This study examines generative AI employment in foreign language teaching preparation, filling a research gap by offering insights into possibilities, advantages, and challenges of this technology. While ChatGPT's proficiency in creating summaries, quizzes, and learning materials could bolster professional development [21], concerns persist regarding reliability and ethical use of AI output in educational contexts where it may affect teaching quality and professional judgment [22]. Therefore, this study investigated generative AI's effects on teachers' preparation experiences and their views on its benefits and limitations, providing knowledge on the applicability and effectiveness of AI use in foreign language teaching preparation. The specific purpose of this study is to empirically investigate, in a single university setting in central China, how a four-week generative AI training intervention influences foreign language teachers' preparation experiences across curriculum design, instructional resource development, teaching activity planning, work motivation, and work efficiency, and to document their perceptions of the strengths, challenges, and recommendations regarding generative AI use. In contrast to prior studies that have primarily examined AI in learner-facing contexts such as student writing evaluation [15] and question generation for English education [14], this study focuses explicitly on teacher-facing preparation workflows across multiple language subjects (English, Japanese, German, French, and Spanish), an angle that has not been systematically examined in the literature.

To comprehensively investigate generative AI's role in foreign language teaching preparation, this study was guided by two research questions. First, how does generative AI implementation affect teachers' work experiences, including work efficiency, professional development, and instructional design capabilities? Second, what strengths, challenges, and recommendations emerge from teacher perspectives on generative AI use for teaching preparation? The findings illuminate benefits and drawbacks of generative AI in foreign language teaching preparation, guiding educational administrators and researchers in developing effective teaching support strategies and training programs. A qualitative case study method was chosen because the phenomenon under investigation, namely teachers' preparation practices with an emerging technology, is process-oriented, context-dependent, and not yet well documented; case study methodology [23], [24] is therefore better suited than a large-scale survey for producing rich, contextualized descriptions of how teachers actually use generative AI and how they interpret its value and limitations in their own words. Scientifically, this study contributes early empirical evidence that can support the design of subsequent larger-scale quantitative and mixed-methods studies on AI-assisted teacher preparation, and it extends the literature on generative AI in education from student-facing applications toward teacher-facing preparation workflows. Socially, the findings can inform the design of teacher training programs and institutional guidelines that help language teachers use generative AI productively while safeguarding professional judgment, academic integrity, and student learning experiences. The remainder of this manuscript is organized as follows: Section II describes the research design, study group, program, interview protocol, data collection, and validity and reliability procedures; Section III presents the results organized by the two research questions; Section IV discusses the findings in relation to prior work, as well as pedagogical and theoretical implications, limitations, and directions for future research; and Section V concludes the paper by summarizing the novelty and implications.

II. Methods

A. Research Design

A qualitative case study design was employed to conduct an in-depth examination of the phenomenon within its actual context and to gain insights into emerging technology applications in teaching practice [23]. Foreign language teachers from different linguistic backgrounds and teaching experience levels were considered as units of analysis. The researchers examined generative AI's effects on teachers' teaching preparation experiences and explored the tool's advantages and drawbacks. Fig. 1 illustrates the overall research design and flow.

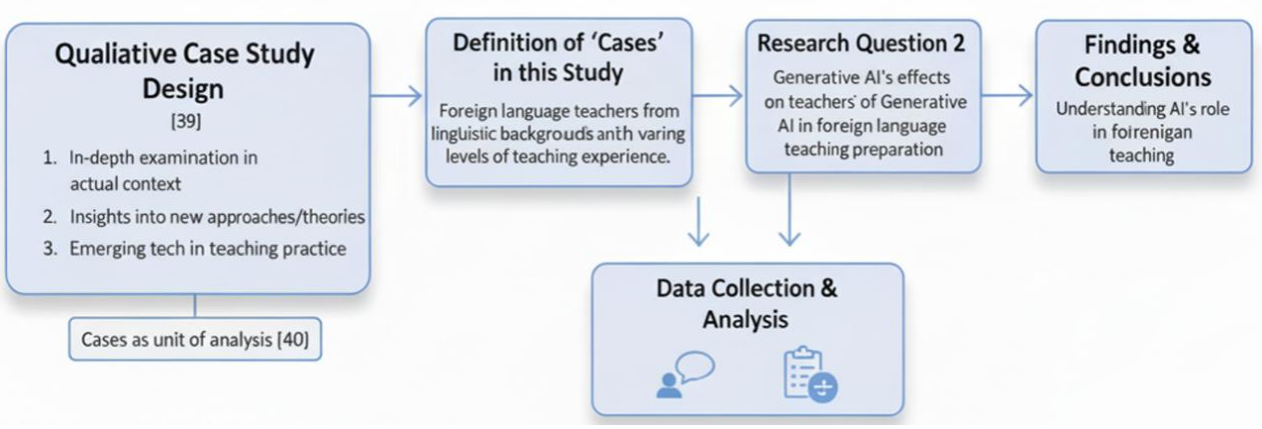


Fig. 1. Research design and flow.

B. Study Group

The study group consisted of 13 foreign language teachers at a comprehensive university in central China. Maximum variation sampling was employed to ensure diversity in linguistic backgrounds and teaching experience levels [23]. Due to administrative concerns about implementing emerging AI technology, the research was conducted with volunteers from a single college. Both training and interviews were conducted online considering teachers' schedules and geographical dispersion. Table I presents the demographic characteristics of the participants.

As shown in Table I, the study group included 10 female and 3 male teachers aged 28-45. Teachers represented diverse linguistic backgrounds including English, Japanese, German, French, and Spanish, with teaching experience ranging from 3 to 20 years. Four participants held doctoral degrees and nine held master's degrees. Most teachers reported using ChatGPT for teaching preparation for the first time during this study.

TABLE I. DEMOGRAPHIC CHARACTERISTICS OF THE PARTICIPANTS

Teacher Code	Gender	Age	Teaching Language	Teaching Experience (years)	Education Level
T1	Female	32	English	8	Master's
T2	Female	29	English	5	Master's
T3	Female	35	Japanese	12	Doctoral
T4	Female	41	English	15	Doctoral
T5	Male	38	English	13	Master's
T6	Male	45	German	20	Doctoral
T7	Female	33	English	9	Master's
T8	Male	36	English	11	Master's
T9	Female	28	French	3	Master's
T10	Female	31	English	7	Master's
T11	Female	39	English	14	Doctoral
T12	Female	42	Spanish	18	Doctoral
T13	Female	30	English	6	Master's

C. Program Description

The program engaged teachers in four 90-minute online training sessions integrating generative AI into curriculum design, instructional resource development, teaching activity planning, and assessment design. Training modules were designed to empower teachers in enhancing their teaching preparation skills and work efficiency through the effective use of generative AI.

D. Training Materials

Training materials included tailored plans, presentation slides, and ChatGPT 3.5 free version. Curriculum design training covered unit topic design, learning objective formulation, lesson flow planning, and materials selection. Teachers used generative AI to generate instructional ideas, create teaching scenarios, and develop materials. Activity design training focused on brainstorming ideas, formulating steps, and obtaining implementation suggestions. Assessment design training involved practicing AI use for generating test items, designing scoring rubrics, and formulating formative assessment schemes.

E. Teaching Preparation Activities

Training plans comprised AI-assisted curriculum design, instructional materials development, teaching activity planning, assessment design, and teaching reflection activities. Fig. 2 shows the various ways teachers engaged with generative AI in different teaching preparation activities. As Fig. 2 illustrates, teachers used the tool to create unit lesson plans, generate learning objectives, and design teaching activities. Fig. 3 depicts the instructional design frameworks that two participants created for "Business English" and "Academic Writing" topics.

As shown in Fig. 3, teachers used the tool to generate discussion questions for group discussions and worked in pairs on AI-assisted teaching reflection exercises. Each teacher received a teaching scenario prompt to use with ChatGPT for brainstorming and planning teaching reflections, which were then shared with partners in breakout rooms for discussion. The facilitator provided standardized prompt templates initially, but teachers were encouraged to write their own prompts in the following weeks to foster independence and creativity. Teachers were guided to reference ChatGPT's answers while incorporating original professional judgment in their teaching preparation work.

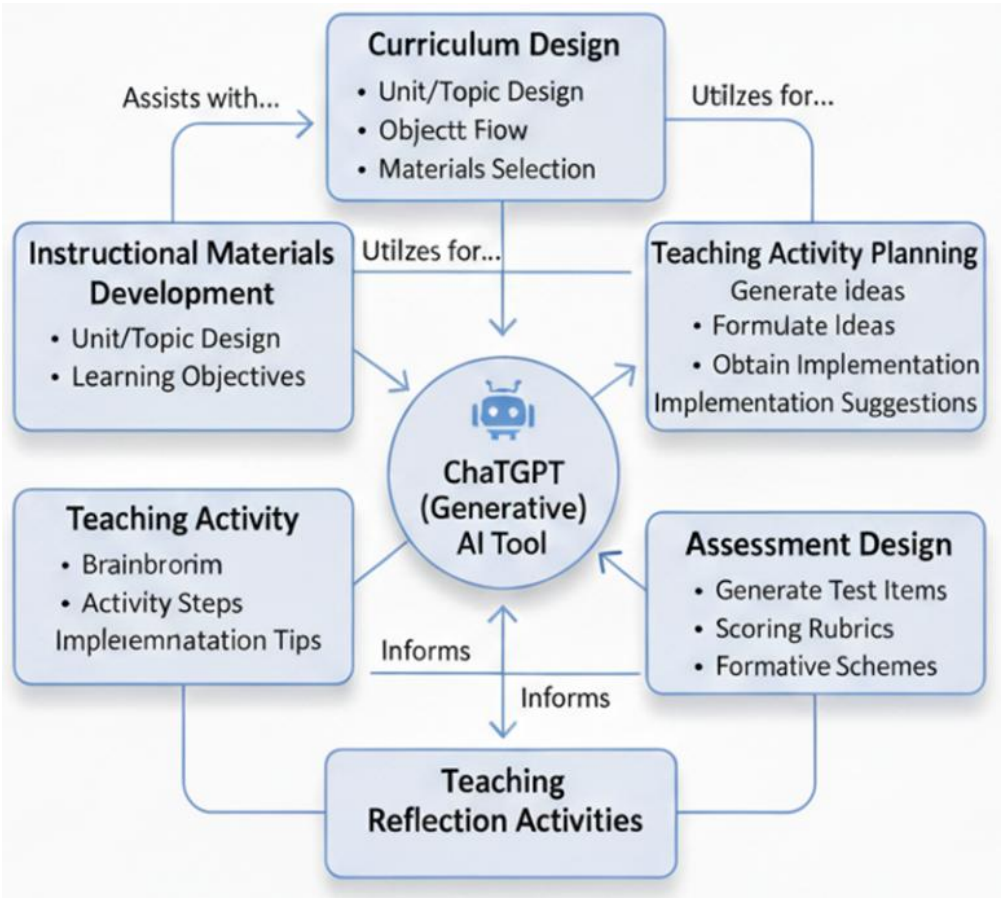


Fig. 2. Integration of ChatGPT in teaching preparation activities.

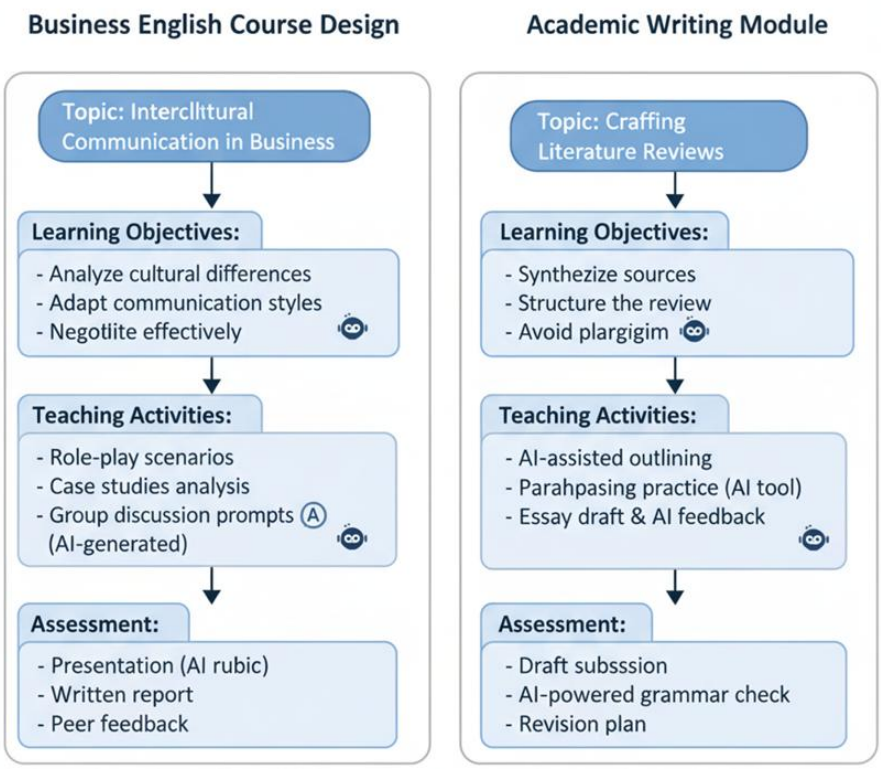


Fig. 3. Examples of instructional designs created with ChatGPT.

F. Data Collection and Analysis Process

The facilitator explained the study's purpose and data confidentiality to teachers and obtained informed consent. Teachers participated in four weeks of training activities followed by semi-structured interviews conducted online via Tencent Meeting platform. Participants indicated time preferences through a questionnaire, and interviews lasting 30-40 minutes were conducted accordingly with audio recording. The scheduling questionnaire used to obtain time preferences collected only nominal-level scheduling data (date and time-slot choices) and was used solely for logistical purposes; it was not used to generate study findings. All substantive study data were collected through open-ended semi-structured interview questions (not closed-ended survey items), and therefore no Likert-type ordinal or interval/ratio measurement scales were involved. The full interview protocol, including all open-ended questions and probes, is available from the corresponding author upon reasonable request, and a read-only version has been deposited at a public repository link that will be provided in the final version of the manuscript. Because the study's substantive instrument is a qualitative interview protocol rather than a quantitative survey, classical psychometric tests of reliability (e.g., Cronbach's alpha) and construct validity are not applicable; instead, trustworthiness was ensured through expert review of the interview guide, inter-coder reliability (simple percentage agreement of 89% and Cohen's Kappa), and maximum variation sampling (see the subsection on Validity and reliability). Given the sample size (n = 13, below the commonly cited n ≥ 30 threshold for quantitative generalization), no inferential statistical tests were performed, and the findings are explicitly framed as institution-specific qualitative evidence rather than as statistically generalizable results.

Semi-structured interview questions were developed within the established case study methodology framework [24], informed by review articles on AI applications in teacher professional development [25]. Interview questions addressed generative AI's applications, impacts, benefits, drawbacks, and optimization suggestions in teaching preparation. Audio-recorded interviews were transcribed verbatim, and three researchers established a coding and category system. Each researcher independently coded the transcriptions, and coding systems were combined after reviewing similarities and differences until consensus was reached.

G. Validity and Reliability

Credibility and transferability were ensured through multiple strategies [26]. Four researchers with foreign language education experience prepared the semi-structured interview form based on training feedback. Five experts reviewed the questions for clarity, resulting in necessary amendments. The maximum variation sampling method secured transferability by examining teachers from various linguistic backgrounds and teaching experience levels. Codes and themes are clearly exhibited with direct quotations from interview data.

Reliability was enhanced through several measures [27]. Interviews were conducted immediately after training sessions to minimize contextual deviation. Researchers posed identical questions to each participant, and all interviews were recorded for verification. Three researchers with doctoral degrees in curriculum and instruction independently coded the data. Inter-coder reliability (ICR) was calculated using the simple percentage agreement formula [28]:

$$\text{Agreement} = \frac{\text{Number of Agreements}}{\text{Total Number of Codes}} \times 100\% \quad (1)$$

Based on this formula, the three researchers achieved a simple percentage agreement of 89%, which is above the 80% threshold commonly reported as acceptable for thematic coding in qualitative research [28]. To further validate the coding consistency, Cohen's Kappa coefficient was also calculated to account for chance agreement:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

where P_o represents the observed agreement proportion and P_e represents the expected agreement by chance. The computed Cohen's Kappa value was 0.83, indicating substantial-to-almost-perfect agreement according to the conventional interpretive ranges. It should be noted that these statistics apply to the inter-coder reliability of the qualitative thematic coding, not to a psychometric survey instrument, since no closed-ended quantitative survey was used in this study. Given that the study group (n = 13) falls below the commonly cited n ≥ 30 threshold for quantitative generalization, the interview instrument is considered sufficient for qualitative case study analysis, but the findings are framed as institution-specific and are not intended as a statistically generalizable reference. As shown in Tables II and III, thematic analysis revealed emergent patterns in teacher perceptions regarding generative AI's impact on teaching preparation; the coded themes were then interpreted in alignment with the two research questions.

III. Results

Thirteen teachers were interviewed to investigate generative AI-assisted teaching preparation's effectiveness on work experiences and views about its benefits and limitations. Anonymity was ensured using code names (T1, T2, T3, etc.) consistently across all tables. Teachers' views were analyzed and organized into themes, presented with detailed findings.

A. The Impact of Generative AI on Teachers' Foreign Language Teaching Preparation Experiences

The findings suggest generative AI positively impacts curriculum design, instructional materials development, and teaching activity planning (T1, T3, T4, T5, T6, T8, T10, T11, T12). Teacher T3 noted that ChatGPT helped learn new teaching strategies in authentic contexts with specific examples for classroom implementation. However, limited impact was observed on assessment design (T4, T12) and teaching research capability (T8).

Generative AI improved teachers' work motivation and professional engagement, contributing to more efficient teaching preparation experiences (T1, T7, T8, T9, T10, T12). Teacher T7 appreciated instant feedback and design suggestions that saved time previously spent consulting reference books. Participants emphasized generative AI's adaptability for interactive curriculum design, creative activity development, and exploring pedagogical approaches (T2, T3, T5, T9, T10).

Teacher T9 mentioned that ChatGPT provided cultural information about different French regions, enriching course content authenticity. Table II presents patterns of teachers' views on generative AI's impact on teaching preparation experiences.

TABLE II. IMPACT OF GENERATIVE AI ON TEACHERS' FOREIGN LANGUAGE TEACHING PREPARATION EXPERIENCES

Themes	Category
Impact on specific teaching preparation tasks	Curriculum design capability enhancement: T5, T6, T8, T12 Instructional resource development improvement: T3, T4, T5, T10, T11, T12 Teaching activity planning advancement: T1, T8, T12 Limited impact on assessment design: T4, T12 No impact on teaching research ability: T8
Work motivation and professional engagement	Enhanced work motivation: T1, T7, T8, T9, T10, T12 Elevated professional engagement: T8, T9 No increase in work motivation: T4, T6
Versatility in teaching preparation activities	Creative teaching activity design: T3, T5 Interactive curriculum planning experiences: T2 Cross-cultural teaching content exploration opportunities: T9
Areas for improvement	Occasional technical issues: T7 Limited contribution to deep instructional design: T4

TABLE III. STRENGTHS, CHALLENGES, AND RECOMMENDATIONS ON GENERATIVE AI FOR FOREIGN LANGUAGE TEACHING PREPARATION

Themes	Category
Strengths	Instructional design capability enhancement (T1, T4, T9, T11) Time-saving and efficient resource acquisition (T1, T3, T4, T13) Comprehensive pedagogical knowledge and support (T1, T2, T10) Cross-cultural teaching content development (T1, T2, T10) Teaching vocabulary and expression expansion (T1, T5, T8, T9, T11) Pedagogical knowledge learning and examples (T3, T4, T6, T7, T8) Gamified teaching and flexible use (T1, T5)
Challenges	Over-reliance and capability deterioration concerns (T1, T2) Connection issues and AI-generated errors (T1, T3) Encouraging shortcuts in teaching preparation (T1, T2, T3) Limited effectiveness for oral teaching design (T11) Potential misuse, emotional responsiveness, and accessibility issues (T3, T5)
Recommendations for improvement	Technical enhancements and accessibility improvement (T1, T4, T5) Integration with authentic teaching contexts and interactivity (T2, T5) Novelty and engagement (T2, T5) Efficiency and time-saving capabilities (T3, T13) Addressing potential misuse and limitations (T2, T3) Improved communication and personalization (T3, T4)

Some teachers noted technical difficulties such as unstable connections or insufficiently accurate responses, and limited assistance in deep-level instructional design tasks (T4, T7). Teacher T4 observed that suggestions for complex grammar teaching sequences were often superficial, lacking consideration of students' cognitive developmental stages. Generative AI demonstrates potential to enhance teaching preparation capabilities while fostering motivation and engagement, though refinement remains essential to mitigate constraints and enhance quality.

B. *Teacher Perspectives on Using Generative AI for Foreign Language Teaching Preparation: Strengths, Challenges, and Recommendations*

Teachers' perspectives underscored positive impacts of generative AI in teaching preparation. Teachers praised enhanced instructional design capabilities, with T9 noting improved organization of instructional content and more creative teaching activities using diversified approaches such as task-based language teaching. Generative AI expedited teaching preparation by quickly generating instructional resources and providing pedagogical suggestions (T1, T3, T4, T13). Teacher T1 noted faster access to instructional information, with ChatGPT providing case backgrounds, discussion questions, and role-play activities that saved preparation time.

The tool provided comprehensive pedagogical knowledge and support (T1, T2, T10) and contributed to cross-cultural teaching content development (T1, T2, T10). Teacher T10 observed that ChatGPT provided cultural information about different teaching contexts and celebration differences across English-speaking countries. Teachers mentioned improvements in teaching vocabulary and pedagogical expressions (T2, T5, T8, T9, T11) and enhanced pedagogical knowledge through practical examples (T3, T4, T6, T7, T8). Teacher T6 noted acquiring professional pedagogical terms such as "scaffolding instruction" and "differentiated instruction". Some teachers found the tool useful for gamified teaching (T1, T5). Table III presents teacher perspectives on strengths, challenges, and recommendations of generative AI in teaching preparation.

Teachers identified challenges and potential drawbacks. Concerns about over-reliance emerged (T1, T2), with T1 expressing worry about deteriorating teaching innovation capability and professional judgment. Teacher T1 noted consciously engaging in independent thinking before using ChatGPT as a verification tool rather than replacement.

Concerns included connection issues and AI-generated errors (T1, T3), with T3 recalling inaccurate Japanese honorific examples requiring careful review. Additional concerns involved encouraging shortcuts in teaching preparation (T1, T2, T3) and limited effectiveness for oral teaching design (T11). Teacher T11 noted that oral communication activity suggestions were formulaic, lacking spontaneity and interactivity.

Teachers highlighted potential misuse in assignment design (T3) and need for improved emotional responsiveness and personalized expressions (T3). Teacher T3 explained that ChatGPT's rational responses lacked humanistic warmth when addressing students' learning anxiety. Accessibility issues due to high traffic were noted (T5). Teachers offered recommendations including technical enhancements and increased accessibility (T1, T4, T5), integration with authentic teaching contexts (T2, T5), and improved communication and personalization (T3, T4). Teacher T2 suggested integrating authentic classroom videos to demonstrate strategy application effects. Teacher T4 proposed ChatGPT remember teaching style and preferences for more personalized suggestions.

IV. Discussion

A. The Impact of Generative AI on Teachers' Foreign Language Teaching Preparation Experiences

Results indicate generative AI positively impacts teaching preparation capabilities in curriculum design, instructional materials development, and teaching activity planning, while increasing teacher motivation and engagement [29]. Research confirms that AI-assisted instructional design improves EFL teachers' curriculum planning, instructional organization, pedagogical coherence, and resource application compared to traditional methods [30], and stimulates authentic pedagogical interactions and productivity [31]. These findings align with our results, indicating generative AI supports teachers' professional development across multiple dimensions. The degree of change observed between traditional preparation (teachers' self-reported prior workflow) and AI-assisted preparation was substantial along several dimensions: for curriculum design and instructional resource development, 10 of 13 teachers (approximately 77%) reported a notable reduction in time required for routine tasks (e.g., generating draft lesson outlines, discussion questions, or case backgrounds) compared with their traditional practice, whereas for deep instructional design tasks (e.g., complex grammar sequencing sensitive to students' cognitive development), teachers reported only marginal benefit over traditional methods, consistent with documented limitations of ChatGPT in identifying deeper pedagogical errors [7]. The novel contribution of the present study, beyond what has been reported in [30], [31], lies in the cross-language, teacher-facing focus (English, Japanese, German, French, and Spanish) and in the explicit separation of tasks where AI substantially outperforms traditional practice (curriculum design, resource development) from tasks where it does not (assessment design, oral-communication activity design, teaching research).

Studies emphasize ChatGPT's potential for personalization and collaboration [32], complementing our observation of increased motivation and engagement. The findings are consistent with previous research supporting ChatGPT's effectiveness in teacher professional development [33], though further investigation is needed. ChatGPT offers comprehensive feedback in instructional reflection [34] and improves instructional design capabilities through AI-generated activities and pedagogical feedback [35]. ChatGPT's potential in creating interactive dialogues and cross-cultural communication tasks [36] is particularly relevant in foreign language teaching. Multiple teachers mentioned ChatGPT's assistance in providing cultural background information and designing culturally sensitive activities, though it lacks face-to-face professional development benefits such as practical exercises in authentic contexts and emotional support.

B. Teacher Perspectives on the Advantages and Disadvantages of Generative AI in Foreign Language Teaching Preparation

ChatGPT improves instructional design capabilities by providing teaching ideas, pedagogical feedback, and introducing instructional structures and vocabulary lists, supported by studies emphasizing AI language models' contribution to teaching vocabulary skills [37]. ChatGPT fosters learner-centered teaching and enhances creativity, problem-solving abilities, and professional vocabulary development [38], with AI-assisted training showing significant improvements compared to traditional instruction [30]. ChatGPT eases preparation by saving time through efficient information provision, lesson plan generation, and materials creation [39]. ChatGPT provides opportunities for cultural and creative learning experiences, aiding critical thinking, communication skills, and intercultural competencies vital in language teaching [40], with participants noting enhanced creativity in generating teaching activities.

Participants raised concerns about possible over-reliance on ChatGPT, aligning with documented literature concerns [9]. Researchers recommend teachers maintain vital roles in teaching preparation and exercise caution in AI integration to avoid over-reliance and ensure balanced approaches incorporating technology and human insights. Concerns include emotional and psychological aspects of technology-facilitated learning [41], potential impediment to critical thinking skills and learner autonomy [42], and ChatGPT's generation of incorrect information, biases, ethical concerns, and privacy issues [43]. Teachers indicated occasional technical problems, such as connection issues [44]. These insights suggest ChatGPT's role should be viewed as complementary to traditional teaching methods, ensuring balanced approaches where teachers maximize efficiency while maintaining professional judgment and creativity.

C. Limitations and Recommendations for Future Research

The study's reliance on 13 teachers from a single institution limits generalizability across different disciplinary backgrounds, teaching experience levels, and educational environments. Future research should employ larger and more diverse sample groups across different institutional types, regions, and teacher backgrounds. Technological access and connectivity issues may have prevented full ChatGPT utilization. Teachers used the free ChatGPT 3.5 version, raising questions about whether premium versions would demonstrate different impacts. Future research could compare pedagogical effects of different ChatGPT

versions or other generative AI tools in teaching preparation. Specifically, because all participants were recruited on a volunteer basis from a single comprehensive university in central China under administrative constraints, the findings should be read as institution-specific qualitative evidence rather than as generalizable claims about generative AI in foreign language teaching broadly. In addition, the tool examined in this study was limited to ChatGPT 3.5 (free version); given the substantial capability differences between GPT-3.5 and subsequent model versions (e.g., GPT-4, GPT-4o, and later releases), the applicability of the present findings to more capable current or future generative AI tools is constrained, and the conclusions should not be read as applying to generative AI broadly. A further limitation is that the study includes no student outcome data and no objective measure of lesson quality; all reported benefits such as improved efficiency, enhanced instructional design, and increased motivation are self-reported teacher perceptions, and without external validation (e.g., classroom observation, student learning gains, or blinded expert ratings of lesson plans) it is not possible to establish whether AI-assisted preparation actually improves instructional effectiveness. Future research should therefore combine teacher self-reports with objective measures of lesson quality and student outcomes, compare model versions (e.g., GPT-4/4o and domain-specific tools), and include multi-institution samples.

Guidelines and protocols for ethical and responsible ChatGPT use in teaching preparation should be developed, clearly defining application scope, usage principles, and preventive measures for potential risks. Future research involving quantitative studies to investigate ChatGPT's effectiveness in influencing teachers' professional development outcomes, job satisfaction, teaching beliefs, and attitudes would support more valid, reliable, and generalizable findings. Adopting mixed research methods would combine qualitative depth insights with quantitative breadth coverage, providing comprehensive perspectives on generative AI's role in teaching preparation.

D. Pedagogical Implications

The study holds significant pedagogical implications for foreign language teachers. Generative AI enables teachers to personalize and manage teaching preparation processes at their own pace, unlike traditional standardized training models. AI tools provide customized suggestions based on specific needs and teaching contexts, helping teachers develop unique teaching styles and professional specializations. The interactive nature of generative AI fosters active professional engagement through curriculum design, instructional resource development, and teaching activity planning. However, ChatGPT has not yet been effective in improving oral teaching skills, whereas other AI applications like "Elsa Speak" benefit pronunciation teaching [45], suggesting the need to select appropriate AI tools for different teaching skills.

Generative AI provides varied content enabling teachers to gain insights into different cultural backgrounds and pedagogical structures, particularly important for cultivating global perspectives and cross-cultural teaching competencies. In globalization contexts, foreign language teachers must cultivate students' intercultural communicative competence, where AI tools offer unique advantages in providing multicultural perspectives. Teachers should incorporate generative AI as supplementary support while engaging in authentic teaching practice reflection. Balancing generative AI use with professional autonomy development prevents possible teaching capability deterioration and over-reliance. Teachers should view AI as enhancing rather than replacing professional judgment, critically evaluating AI-generated content and adapting it according to specific teaching contexts while maintaining sensitivity to student needs.

E. Theoretical Implications

Generative AI integration into teaching preparation presents theoretical implications for educational theory. Generative AI demonstrates compatibility with personalized learning approaches [46], offering individualized feedback and guidance tailored to individual teaching needs. Unlike traditional "one-size-fits-all" training models, AI-supported personalized learning provides customized support based on knowledge background, teaching experience, and professional development needs, reflected in content, learning pace, and depth.

Aligned with Self-Regulated Learning theory [47], generative AI guides teachers in teaching strategy selection and time management. Teachers learn through AI interaction to set teaching preparation goals, monitor progress, evaluate outcomes, and make adjustments. This self-regulation capability cultivation is crucial for long-term professional development, enabling teachers to become active managers of professional growth. Within Cognitive Load Theory [48], generative AI's adaptable nature optimizes cognitive processing by adjusting information complexity based on teachers' knowledge level and comprehension ability, avoiding cognitive overload or resource waste. Under Engagement Theory [49], generative AI's interactive nature enhances teachers' active participation through prompt feedback and interactive learning experiences, promoting deep cognitive engagement and emotional investment that maintain sustained learning motivation and cognitive processing. These theoretical lenses are consistent with recent empirical work on ChatGPT in language education: prior IJACSA studies have demonstrated that ChatGPT can generate chatbot dialogues suitable for English-as-a-foreign-language learners at the CEFR A2 level [50] and can provide effective personalized learning experiences in higher education when guided by established instructional models such as the 5Es framework [51], offering complementary empirical support for the theoretical mechanisms discussed above.

V. Conclusion

In conclusion, we explored the impact of generative AI-assisted foreign language teaching preparation experiences through the views of university foreign language teachers on the opportunities and challenges of generative AI-assisted teaching preparation. The findings revealed generative AI's positive impact on curriculum design and limited impact on teaching research, development of pedagogical knowledge and instructional resource application capabilities, motivation of participants' professional development, and promotion of interactive teaching preparation experiences. Moreover, generative AI was found to contribute to participants' understanding of diverse teaching cultures and fostered their creative thinking capabilities during training activities.

However, a significant concern that emerged from the results was the potential over-reliance on generative AI. The novelty of this study lies in three aspects. First, it provides cross-language empirical evidence on teacher-facing (not student-facing) preparation workflows with generative AI across five languages (English, Japanese, German, French, and Spanish), whereas prior studies have predominantly examined AI in student-facing tasks such as writing evaluation or question generation. Second, it differentiates the magnitude of AI's impact across specific preparation sub-tasks, showing that curriculum design, instructional resource development, and teaching activity planning benefit substantially (approximately 62–77% of teachers reporting positive impact), whereas assessment design, oral-teaching design, and teaching research benefit only marginally. Third, it documents a concrete pattern of over-reliance risk that emerges even within a short four-week intervention, thereby informing institutional guidelines before large-scale adoption. The implications for science are that future research should move beyond self-reported teacher perceptions toward externally validated measures of lesson quality and student outcomes, and should systematically compare generative AI model versions. The implications for society are that universities and teacher-training programs should treat generative AI as a complement to, not a substitute for, professional judgment, and should develop ethical-use guidelines that explicitly define application scope, prevent over-reliance, and protect academic integrity. The main shortcomings of the present study, namely the small single-institution sample (n = 13), the exclusive use of ChatGPT 3.5, and the absence of objective lesson-quality and student-outcome measures, should be regarded as directions for further research: future work should employ larger multi-institution samples, compare more capable model versions (e.g., GPT-4/4o and successor models), and combine teacher self-reports with classroom observation and student learning data.

ACKNOWLEDGMENTS

This work was supported by the Soft Science Research Project of Henan Province (Grant No. 252400410584) and the Higher Education Research Project of Henan Province (Grant No. 2025SXHLX030).

REFERENCES

[1] OpenAI. ChatGPT (GPT-4, September version) [Multimodal Large Language Model]. 2024. Retrieved January 02, 2024. https://openai.com/chatgpt

[2] Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228-239.

[3] Shen, Y., Heacock, L., Elias, J., Hentel, K., Reig, B., Shih, G., Moy, L. (2023). ChatGPT and other large language models are double-edged swords. *Radiology*, 307(2), 1-4.

[4] Ouyang, L., Wu, J., Xu, J., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.

[5] Gunawan, J. (2023). Exploring the future of nursing: Insights from the ChatGPT model. *Belitung Nursing Journal*, 9(1), 1-5.

[6] Moorhouse, B. L. (2024). Beginning and first-year language teachers' readiness for the generative AI age. *Computers and Education: Artificial Intelligence*, 6, 100201.

[7] Algaraady, J., & Mahyoob, M. (2023). ChatGPT's capabilities in spotting and analyzing writing errors experienced by EFL learners. *Arab World English Journal*, 9, 3-17.

[8] Bao, Y., & Li, B. (2023). A preliminary study on graduate student instructors' exploration, perception, and use of ChatGPT. *International Journal of Computer-assisted Language Learning and Teaching*, 13(1), 1-23.

[9] Mohamed, A. M. (2024). Exploring the potential of an AI-based Chatbot (ChatGPT) in enhancing English as a Foreign Language (EFL) teaching: Perceptions of EFL Faculty members. *Education and Information Technologies*, 29, 3195-3271.

[10] Kim, S., & Lee, Y. (2024). Investigation into the influence of socio-cultural factors on attitudes toward artificial intelligence. *Education and Information Technologies*, 29, 9907-9935.

[11] Bauer, E., Greisel, M., Kuznetsov, I., Berndt, M., Kollar, I., Dresel, M., Fischer, M. R., & Fischer, F. (2023). Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. *British Journal of Educational Technology*, 54(5), 1222-1245.

[12] Kim, J., Lee, H., & Cho, Y. H. (2022). Learning design to support student-AI collaboration: Perspectives of leading teachers for AI in education. *Education and Information Technologies*, 27(5), 6069-6104.

[13] Ayotunde, O. O., Jamil, D. I., & Çavuş, N. (2023). The impact of artificial intelligence in foreign language learning using learning management systems: a systematic literature review. *Information Technologies and Learning Tools*, 95(3), 215-228.

[14] Lee, U., Jung, H. C., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2024). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies*, 29, 11483-11515.

[15] Yang, H., Gao, C., & Shen, H. (2024). Learner interaction with, and response to, AI-programmed automated writing evaluation feedback in EFL writing: An exploratory study. *Education and Information Technologies*, 29, 3837-3858.

[16] Zaki, N., Turaev, S., Shuaib, K., Krishnan, A., & Mohamed, E. (2023). Automating the mapping of course learning outcomes to program learning outcomes using natural language processing for accurate educational program evaluation. *Education and Information Technologies*, 28(12), 16723-16742.

[17] Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57, 100745.

[18] Tavares, C., Oliveira, L., Duarte, P., & Da Silva, M. M. (2023). Artificial intelligence: A blessing or a threat for language service providers in Portugal. *Informatics (Basel)*, 10(4), 81.

[19] Alghasab, M. B. (2025). English as a foreign language (EFL) secondary school students' use of artificial intelligence (AI) tools for developing writing skills: unveiling practices and perceptions. *Cogent Education*, 12(1), 1-10.

[20] Pack, A., & Maloney, J. (2023). Using generative artificial intelligence for language education research: Insights from using OpenAI's ChatGPT. *TESOL Quarterly*, 57(4), 1571-1582.

[21] Khan, R. A., Jawaid, M., Khan, A. R., & Sajjad, M. (2023). ChatGPT - reshaping medical education and clinical management. *Pakistan Journal of Medical Sciences*, 39(2), 605-607.

[22] Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*, 20(2), 1-26.

[23] Patton, M. Q. (2022). Qualitative research and evaluation methods. Sage.

[24] Yin, R. K. (2011). Applications of case study research. Sage.

[25] Yang, H., & Kyun, S. (2022). The current research trend of artificial intelligence in language learning: A systematic empirical literature review from an activity theory perspective. *Australasian Journal of Educational Technology,* 38(5), 180-210.

[26] Lucas, J. W. (2003). Theory-testing, generalization, and the problem of external validity. *Sociological Theory*, 21(3), 236-253.

[27] Cohen, L., Manion, L., & Morrison, K. (2017). Research Methods in Education. 8th ed. Routledge.

[28] Miles, M., Huberman, A. M., & Saldaña, J. (2019). Qualitative data analysis: A methods sourcebook. Sage Publications.

[29] Shadiev, R., & Liang, Q. (2023). A review of research on AR-supported language learning. *Innovation in Language Learning and Teaching*, 18(1), 78-100.

[30] Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, 14, 1260843.

[31] Nugroho, A., Putro, N. H. P. S, & Syamsi, K. (2023). The potentials of ChatGPT for language learning: Unpacking its benefits and limitations. *Register Journal*, 16(2), 224-247.

[32] Klayklung, P., Chocksathaporn, P., Limna, P., Kraiwanit, T., & Jangjarat, K. (2023). Revolutionizing education with ChatGPT: Enhancing learning through conversational AI. *Universal Journal of Educational Research*, 2(3), 217-225.

[33] Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, 28(11), 13943-13967.

[34] Guo, K., & Wang, D. To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, 29, 8435-8463.

[35] Hwang, W. Y., Nurtantyana, R., Purba, S. W. D., Hariyanti, U., Indrihapsari, Y., & Surjono, H. D. (2023). AI and Recognition Technologies to facilitate English as Foreign Language writing for supporting personalization and contextualization in authentic contexts. *Journal of Educational Computing Research*, 61(5), 1008-1035.

[36] Topsakal, O., & Topsakal, E. (2022). Framework for a foreign language teaching software for children utilizing AR, voicebots and ChatGPT (large language models). *The Journal of Cognitive Systems*, 7(2), 33-38.

[37] Akopiants, N. (2023). Using ChatGPT in the process of learning English: Advantages and opportunities. *Actual Problems of Ukrainian Society Development*, 1, 69-72.

[38] Rospigliosi, P. (2023). Artificial intelligence in teaching and learning: What questions should we ask of ChatGPT? *Interactive Learning Environments*, 31(1), 1-3.

[39] Stepanechko, O., & Kozub, L. (2023). English teachers' concerns about the ethical use of ChatGPT by university students. *Grail of Science*, 25, 297-302.

[40] Michalon, B., & Camacho-Zúñiga, C. (2023). ChatGPT, a brand-new tool to strengthen timeless competencies. *Frontiers in Education*, 8, 1251163.

[41] Sampson, R. J., & Yoshida, R. (2020). L2 feelings through interaction in a Japanese-English online chat exchange. *Innovation in Language Learning and Teaching*, 15(2), 131-142.

[42] Sun, G., & Hoelscher, S. H. (2023). The CHATGPT storm and what faculty can do. *Nurse Educator*, 48(3), 119-124.

[43] Mohammad, B., Supti, T., Alzubaidi, M., Shah, H., Alam, T., Shah, Z., & Househ, M. (2023). The Pros and Cons of Using ChatGPT in Medical Education: A Scoping Review. *Stud. Health Technol. Inform*, 305, 644-647.

[44] Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning-are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), 237-257.

[45] Akhmad, N. W., & Munawir, A. (2022). Improving the students' pronunciation ability by using Elsa Speak App. Ideas: *Journal on English Language and Learning Linguistics and Literature*, 10(1), 846-857.

[46] Walkington, C., & Bernacki, M. L. (2020). Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. *Journal of Research on Technology in Education*, 52(3), 235-252.

[47] Zimmerman, B. J., & Pons, M. M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. American Educational Research Journal, 23(4), 614.

[48] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285.

[49] Kearsley, G., & Shneiderman, B. (1998). Engagement theory: A framework for technology-based teaching and learning. *Educational Technology*, 38(5), 20-23.

[50] Young, J. C., & Shishido, M. (2023). Investigating OpenAI's ChatGPT potentials in generating chatbot's dialogue for English as a foreign language learning. *International Journal of Advanced Computer Science and Applications*, 14(6): 65-72.

[51] Albdrani, R. N., & Al-Shargabi, A. A. (2023). Investigating the effectiveness of ChatGPT for providing personalized learning experience: a case study. *International Journal of Advanced Computer Science and Applications*, 14(11):1208-1213.

Climate Change on Social Media: AI and Deep Learning-Based Analysis of Tweets

Bahar URHAN[1], Mehmet KAYAKUŞ[2], Dilşad ERDOĞAN[3],
Gülten ADALI[4]*, Emrah BOZKURT[5], Zeynep Nihan BAKIR[6]
Department of Public Relations and Promotion-Faculty of Communication, Akdeniz University, Antalya 07058, Türkiye[1]
Department of Biomedical Engineering-Faculty of Engineering, Akdeniz University, Antalya 07058, Türkiye[2]
Department of Finance, Banking and Insurance-Korkuteli Vocational School, Akdeniz University, Antalya 07800, Türkiye[3]
Department of Advertising-Faculty of Communication, Akdeniz University, Antalya 07058, Türkiye[4]
Department of Public Relations and Publicity, Afyon Kocatepe University, Afyonkarahisar 03030, Türkiye[5]
Independent Researcher, Muğla, 48000, Türkiye[6]

***Abstract*—The study analyses Turkish and English tweets about climate change on the social media platform Twitter and comparatively examines individuals" perceptions, concerns, and emotional reactions to this issue. A total of 2,046 Turkish and 18,000 English tweets were collected; 1,104 Turkish and 6,449 English tweets were analyzed after the cleaning process. Artificial intelligence-based methods such as text mining, sentiment analysis, and topic modelling are used. Topic modelling with Latent Dirichlet Allocation (LDA) identified prominent themes in tweets in both languages. Sentiment analysis is performed using deep learning techniques to categories tweets into positive, negative, and neutral categories. The findings show that English tweets contain stronger emotional reactions, while Turkish tweets contain a higher proportion of neutral expressions. Additionally, it was observed that the perception of climate change can differ in local and global contexts. Based on a multidimensional analysis of social media data, the study provides valuable insights into the development of environmental communication strategies. The comparison of Turkish and English tweets contributes to understanding the effects of cultural contexts on climate change perception. The findings have important implications for policymakers and environmental awareness campaigns, as they highlight the need for tailored communication strategies that consider cultural differences in climate change perception.**

***Keywords*—Climate change; awareness; social media; communication; machine learning; sentiment analysis; deep learning**

I. Introduction

Climate change is recognized as one of the greatest environmental, economic, and social challenges of our time. Scientific evidence shows that this phenomenon significantly threatens not only natural systems but also human life and global economic balances [1]. In the 21st century, accelerating climate change due to increasing carbon emissions and greenhouse gases has made social awareness and collective action more important than ever. However, the solution to such a complex and multidimensional problem is closely related not only to scientific developments but also to the attitudes and behaviors of societies shaped through access to information and communication channels. In this context, social media platforms play a central role in educating and raising awareness about climate change.

This study aims to examine how Turkish and English tweets on the social media platform Twitter create a social perception in the context of climate change and how users feel about this issue. The study is based on a dataset consisting of 2046 Turkish and 18,000 English tweets, and after the data cleaning process, 1104 and 6449 tweets were analyzed, respectively. The methods used include text mining, sentiment analysis, and topic modelling, and thus, social media content in different languages is evaluated from a comparative perspective.

In this study, text mining and deep learning methods are used to analyses social media data. Texts were cleaned with natural language processing (NLP) techniques, and n-gram analyses were performed. Latent Dirichlet Allocation (LDA), an unsupervised machine learning method, is used to identify themes in the tweets, allowing for a comparison of themes in different languages. In the sentiment analysis phase, Long Short-Term Memory (LSTM) with a deep learning model is used to classify tweets as positive, negative and neutral. These methods reveal the climate change perceptions and emotional tendencies of social media users in a multidimensional way.

The originality of the study stems from the fact that it comparatively analyses how a global problem such as climate change is perceived at local and global levels through social media data. Although there are many studies in the literature on the impact of social media on environmental awareness, such cross-lingual comparisons are limited. For example, Lewandovsky et al. (2019), while examining how social media affects the perception of climate change, stated that analyses in the local and global contexts are limited [2]. This study aims to fill this gap by revealing how perceptions of climate change differ in different cultural and linguistic contexts through Turkish and English tweets.

Unlike recent multilingual studies that rely on transformer-based architectures such as BERT or XLM-R for cross-lingual alignment, this study adopts a comparative approach using language-specific datasets without explicit embedding alignment. This allows for the exploration of natural linguistic divergence in sentiment and topic structures.

Based on this motivation, the study tests the following hypothesis:

*Corresponding author.

H1: There are statistically significant differences between Turkish and English tweets in terms of sentiment distribution and topic structure.

This research makes three main contributions to literature: Firstly, by providing a multidimensional analysis of social media content on climate change, it facilitates the understanding of individuals' perceptions and attitudes towards the issue. Secondly, the comparative evaluation of social media content in different languages using sentiment analysis and topic modelling methods reveals the effects of cultural contexts. Thirdly, the findings have the potential to provide concrete recommendations for policymakers and stakeholders developing environmental communication strategies.

As the impacts of climate change are rapidly increasing, it is critical to understand individuals' perceptions on this issue and develop the right communication strategies. The results of this study may contribute to raising awareness on combating climate change and shed light on the development of more effective environmental communication strategies through social media.

II. Literature

The last two centuries, during which technology has made it possible to create considerably accurate measures, have not only revealed the phenomenon of climate change but also given us a better understanding of the threat it poses to the well-being of life on Earth and the ecological stability of the planet. These measures have led to a growing international recognition of the threat. In the 20th century, most scientists have not seen greenhouse warming as a severe problem. It took decades of evidence, fierce debates, and hundreds of panels in many countries to convince them they were wrong. In the late 1970s, these efforts led to the conclusion that warming might one day become a problem. But, in the current century, scientific meetings and respected panels have concluded that warming is likely to become a serious problem [3].

Considering current knowledge, climate change is recognized as one of the greatest threats to human life. However, its trajectory can be somewhat affected depending on human activities in the face of this threat [1]. Climate change is predicted to seriously impair the life support systems of many species, drastically reduce the population, and result in crucial and fundamental changes to social structures. Therefore, the importance of communication on the topic is revealed by how to approach it and guarantee cooperation, creative policies, appropriate technological developments, and new behaviors [4]. Pearce et al. (2019), social media such as Twitter are important tools for enabling everyday debate and negotiation on climate change [5]. Dutton argues that the internet acts as a "Fifth Estate" and uses estate theory to illuminate the most significant power shift of the digital age [6]. Through the power of the internet, networked individuals can control facts and make social interventions in a civic context. The internet gives them this power, and digital media essentially increases the information and communication power of ordinary individuals. Communities that can independently and strategically search, create content, network, and collaborate on any topic are able to leak information in this way.

The relationship between social media and the construction of people's ideas about climate change and changing attitudes has been the focus of many studies. For example, Lewandovsky et al. (2019) measured attitudes towards climate change on social media and found that science-based posts reinforced belief in global climate change, but negative comments on these posts reduced perceived consensus [2]. It was observed that social media is a more powerful tool in terms of receiving information and forming attitudes with this information. Accordingly, new media and user-generated content have twice the impact size compared to traditional and centralized content-generating channels [7].

Comparative studies on the relationship between media coverage and climate change, such as our current study, are limited in number but have been able to provide estimates with quite significant data. For example, Thaker's (2023) large-scale comparative study on the coverage of climate change in international media, based on data from the 2022 Yale Meta survey of 110 countries, found that the relationship between perceptions formed through social media and protest intention is related to personal harm. Accordingly, because poor and developing countries are more exposed to the impacts of climate change, they perceive climate change as a greater risk and are more willing to participate in protests, even if such content receives low media coverage [8].

The nature of the media helps shape public opinion on climate change [1]. By their very nature, online posts have several advantages over traditional media in influencing public opinion. Digital activism, which stands out among these, provides an advantage. The most prominent example of this is the case of Greta Thunberg. As seen in the phenomenon of raising online awareness through digital activism, which has entered the literature as the "Greta Effect," Greta Thunberg has influenced both policymakers and the news media to take responsibility for climate change, thanks to her supporters or critics [9].

The existing literature on social media assumes that three main mechanisms work in climate change through posts: norm diffusion, opinion leadership, and public opinion formation [10]. On the other hand, research has found that those who express an opinion on this issue exhibit some categorical reactions: deniers, jokers, advocates, skeptics, and realists [11-14]. Considering such categories in climate communication can better determine its direction. Yet another study found that a high concern for climate resulted in greater engagement in online climate change, and adolescents who cared more about the norms of their peers had higher levels of online climate change because of incidental exposure [15]. In this context, it is seen that many variables, approaches, and categories can be revealed through social media research on climate change. On the other hand, due to the rapid advancement in communication technologies, climate-based information disinformation has become even more possible, and misinformation spread has become extremely easy and fast [2]. Therefore, the fact that a very useful and slippery ground such as social media plays an active role in climate communication emphasizes conscious and effective use as well as some precautions.

Just as Climate Central, which "conveys the science, impacts, and solutions of climate change to the public and decision-makers" (2024), provides reliable information on the issue through its posts on various media channels, supporting social media content with realistic arguments such as visuals and videos, the active participation of scientists, who are seen to have a significant impact on public participation and the responsibility of providing first-hand information, and the role of scientists being free from doubt among both the public and their colleagues can make dialogic and participatory climate communication possible [12].

III. Materials and Methods

In this study, data were collected from the Twitter platform in Turkish and English to analyses social media posts on climate change. The data was collected using the API of the X platform over a period of one month. In total, 2046 Turkish and 18,000 English tweets were obtained. During the data collection process, the keyword "climate change" is used. The focus of the study is to understand the perceptions, concerns, and perspectives of users on this issue. In the text preprocessing stage, repetitive, incomplete, noisy, and irrelevant content was removed, and finally, 1104 Turkish and 6449 English tweets were included in the study.

Formally, the sentiment classification problem can be defined as follows: given a tweet x_i, the goal is to predict its sentiment label $y_i \in$ {positive, negative, neutral}. The model aims to learn a function $f(x_i) \rightarrow y_i$ that minimizes classification error.

The LSTM model was implemented with the following configuration: embedding dimension = 100, batch size = 32, epochs = 10, optimizer = Adam, and dropout rate = 0.5. All experiments were conducted using TensorFlow/Keras, and model evaluation is performed using a 70/30 train-test split.

Text mining, topic modelling, and sentiment analysis methods are used to analyses the tweets. Text mining enabled the analysis of word frequencies and text patterns in tweets, and thus meaningful information was extracted from the data. Topic modelling is used to identify the prominent themes in tweets and to determine which topics users focus on. Sentiment analysis is used to identify positive, negative, and neutral emotional tones in tweets. These methods enabled comparative analysis of social media content in different languages and made it possible to evaluate perceptions on climate change in a multidimensional way.

A. Data Set

The data set used in the study was obtained through the API provided by the X platform. The main reason for choosing the X platform is that it is a social media platform that is heavily used by users and provides data from a wide user base on social issues such as climate change [16]. This increased the inclusiveness of the study and the generalizability of its results.

The keyword "climate change" was preferred in the data collection process. The reason for this is that climate change is of increasing importance in both global and local contexts and that social media users have various opinions and perceptions on this issue, providing a suitable data source for the analyses targeted by the study. In this context, the choice of keywords aims to facilitate the understanding of users' perceptions and perspectives on the issue.

The data was collected over a period of one month using the Python programming language and the API of the X platform. The reason for choosing a one-month time is that this period provides the intensity and diversity of users' posts on the topic that can form a meaningful sample. In addition, this period ensured that the data set was kept in analyzable sizes and the data processing process was manageable [17, 18]. However, the dataset is limited to a one-month time frame and does not include demographic or geographic metadata due to API constraints. Therefore, it may not fully represent the broader population. Future studies should incorporate longer temporal coverage and user-level metadata to improve representativeness.

In this process, a total of 2,046 Turkish tweets and 18,000 English tweets were collected. During the preprocessing phase, the dataset was systematically refined by removing duplicate entries, noise, incomplete and inconsistent records, as well as irrelevant content, to enhance data quality, reduce potential bias, and ensure the robustness and reliability of subsequent analyses. These procedures are aimed at increasing the accuracy of the analysis process and to obtain meaningful results. After the cleaning process, the number of tweets suitable for analysis was determined to be 1104 for Turkish and 6449 for English. The data cleaning process involved removing duplicate tweets, non-relevant content (such as advertisements), tweets containing only URLs or hashtags, and posts with insufficient textual information. Additionally, tweets with excessive noise, including random characters and non-linguistic symbols, were excluded.

To minimize potential bias, the same cleaning criteria were applied consistently across both datasets. Furthermore, sentiment distributions were checked before and after preprocessing, and no significant distortion in class proportions was observed, suggesting that the filtering process did not systematically bias the results.

Table I presents examples from the dataset, including multilingual tweets addressing climate change issues.

TABLE I. Sample Dataset

Tweet	Language
Whatever the cause, climate change seems to be an area where countries will work together and struggle to maintain a humane life through joint decisions.	Tr
We must stop the development of agricultural land. Global warming and climate change will make it extremely difficult for people to access healthy food and water.	Tr
In recent years, dam occupancy rates have started to decrease due to reasons such as climate change, drought and increase in water use.	Tr
Serious policies should be developed on climate change and drought. Yes, to green policies for a more loveable world...	Tr
Our planet is the only place where humanity lives and therefore its protection is of utmost importance. However, human activities have resulted in several environmental problems such as overconsumption of natural resources, environmental pollution, climate change and habitat loss.	Tr
The climate crisis forces us to reconsider our values.	En

Humanity must fight climate change much harder, or losses and damage, and suffering, will continue to mount.	En
There's no solution to climate change - and no one can do it alone	En
Climate change is the threat that will unite us all.	En
Scientists revive viruses frozen thousands of years ago in the Arctic to predict the dangers that could be unleashed by melting because of climate change	En

B. Text Mining

Text mining is a big data analysis method that generates meaningful information from text data. Text mining is a data mining study whose data source is text, and it enables the transformation of unstructured data into structured data. Text mining is the process of extracting confidential information from the data in the text that does not have a clear format and formatting the disorganized data [19].

The first thing to do with the data in the analysis process is to pre-process the data. In a data mining process, previous studies have shown that the preprocessing process covers up to 60% of the necessary work, and preprocessing the data contributes 75% to 90% to the success of the data mining project. Non-textual characters such as !, ?, /, +, and emoji in the documents are removed from the data set. The stop words that occur in large numbers in the documents are extracted based on the previously prepared stop word list. Since the use of ineffective words during indexing reduces the size of the index files, it increases memory usage and query processing efficiency [20].

N-gram is a method used to search and compare data or to determine the number of repetitions. Term frequency is used to calculate the term weights in a document. Reverse term frequency is used to find the number of occurrences of the word in more than one document and tries to understand whether this word is a term or a conjunction, etc. (stop-words). It is found by taking the absolute value of the logarithm of the number of documents in which the term occurs divided by the number of documents [21].

These feature extraction methods contribute to model performance by enhancing semantic representation of textual data. However, alternative feature representations such as contextual embeddings (e.g., BERT) were not evaluated, which constitutes a limitation of the study.

C. Sentiment Analysis

To train the supervised LSTM model, sentiment labels were generated using a semi-automated approach. Initially, tweets were labelled using lexicon-based sentiment resources and existing benchmark datasets. Subsequently, a subset of the data was manually reviewed to ensure consistency and correctness of the assigned labels. This approach improves the reliability of the training data and reduces potential bias arising from fully automated labelling. Sentiment analysis is the field of study that analyses people's opinions, emotions, evaluations, and attitudes through written language. Users' thoughts about a particular event are extracted through text analysis, and then emotions are predicted based on the analysis. Due to the increasing data flow, sentiment analysis has become one of the most active topics of natural language processing and text mining [22].

Sentiment analysis studies are questioned and analyzed whether the texts have positive, negative, or neutral content. According to the results of this analysis, the attitude of individuals or a certain group on the subject related to the study is determined [23].

The methods used in detecting sentiment polarity fall under two main categories. These categories are machine learning-based methods and word-based methods. In machine learning-based methods, the system is trained with pre-labelled training data, and emotion classification is performed with the trained system. Any of the machine learning classification algorithms can be used. Under this category, methods based on syntactic analysis of sentences using natural language processing methods and tools are used. There is no need for labelled training data as in supervised machine learning methods. With natural language processing tools and methods, sentences are analyzed, and semantic inferences are made by detecting emotion terms in sentences. A dictionary of emotion terms is often used to detect emotion expressions in sentences [24, 25].

D. Machine Learning

Machine learning is an important subfield of artificial intelligence. Its main purpose is to use computational methods to extract knowledge from data. Machine learning has a wide range of applications, including handwriting and speech recognition, robotics and computer games, natural language processing, brain-machine interfaces, etc. [26].

The classification method, one of the basic methods of data mining, is based on the learning algorithm [27]. Classification algorithms divide the data into certain groups (classes) according to their common characteristics by targeting the desired information. Classification algorithms are run in two stages. Firstly, the classification model is created by analyzing the data set determined as training data. In the second stage, the obtained classification model is applied to a new dataset, and the presence of the determined classes in the data is investigated [28].

Deep learning is a subset of machine learning that enables models to learn hierarchical representations from data through multiple layers of neural networks [29]. Unlike traditional machine learning techniques that rely on handcrafted features, deep learning models automatically extract relevant features from raw text data. In the context of sentiment analysis, deep learning provides a powerful framework for capturing complex language patterns and contextual meanings in social media posts [30].

One of the primary advantages of deep learning for text analysis is its ability to process sequential data effectively. Social media posts, such as tweets, are inherently sequential and often contain nuanced expressions, abbreviations, and informal language structures [31]. Deep learning models, particularly those based on recurrent neural networks (RNNs) and transformer architectures, have demonstrated superior performance in sentiment classification tasks compared to traditional machine learning methods [32].

Long Short-Term Memory (LSTM) networks stand out as a specialized type of recurrent neural network (RNN) particularly well-suited for processing sequential data [33]. Unlike

traditional RNNs, which struggle with long-range dependencies due to the vanishing gradient problem, LSTM networks incorporate memory cells and gating mechanisms to retain relevant information over extended sequences. This capability makes LSTM an ideal choice for analyzing textual data, where contextual relationships between words play a crucial role in sentiment classification.

The architecture of an LSTM network consists of three key gates [34]:

- Forget Gate: Determines which information from the previous time step should be discarded.
- Input Gate: Controls how much new information should be added to the memory cell.
- Output Gate: Regulates the amount of stored information that contributes to the current output.

For this study, the LSTM-based sentiment analysis model was trained on a dataset of Turkish and English tweets related to climate change. The preprocessing pipeline involved tokenization, stop-word removal, and word embedding using pre-trained word vectors. The model was structured with multiple LSTM layers, dropout regularization to prevent overfitting, and a SoftMax activation function in the output layer to classify tweets into three sentiment categories.

The training process involved optimizing the categorical cross-entropy loss function using the Adam optimizer. The dataset was split into training and validation sets to evaluate model performance [35]. Additionally, performance metrics such as accuracy, precision, recall, and F1-score were computed to assess the effectiveness of the LSTM model in sentiment classification.

The deep learning models were implemented using TensorFlow and Keras libraries. The input text data was vectorized using word embeddings such as Word2Vec and GloVe to enhance semantic understanding. Hyperparameters, including batch size, learning rate, and the number of LSTM units, were fine-tuned through grid search and cross-validation [36].

To ensure robustness, the model was trained on a diverse dataset comprising tweets in both Turkish and English. The bilingual nature of the dataset posed challenges in terms of language-specific features; however, LSTM's ability to capture sequential dependencies allowed for effective sentiment classification across languages. Data augmentation techniques were also employed to balance class distributions and improve generalization [37].

By leveraging deep learning techniques, particularly LSTM networks, this study effectively classified climate change-related tweets based on sentiment polarity. The proposed approach outperformed traditional machine learning methods by capturing complex linguistic patterns and long-term dependencies in textual data. The insights derived from this analysis provide valuable contributions to understanding public perceptions of climate change and developing targeted environmental communication strategies. It should be noted that this study does not employ explicit cross-lingual embedding alignment or normalization techniques. Instead, Turkish and English datasets are analyzed independently to preserve language-specific characteristics. While this approach limits direct semantic comparability, it enables the identification of natural linguistic differences between languages.

E. Topic Modelling

Topic modelling (TM) is one of the subfields of text mining. Its main goal is to reveal hidden or open topics in documents. It has an important place among the subfields of text mining, especially with its increasing importance and studies carried out in recent years. KM is a method in which subject classification is made unsupervised in a document [38].

Topic modelling methods are based on the principle that topics with probability distributions on words come together randomly and form documents. Latent Dirichlet Allocation (LDA) is a generative graphical model used to model discrete data such as documents and to reveal the topics that make up the document. "Latent" here refers to finding the meaning of the document by discovering the hidden topics that make up the document. What is meant by generative model is the generation of the words in the document by a simple probabilistic process within the framework of hidden (random) variables, i.e., the creation of the document. GDA, which is a completely unsupervised method, does not require any prior knowledge and works based on a bag-of-words approach. While the placement of words in the document is ignored, the co-occurrence of words is used in this method [39].

Each document consists of a random mixture of topics, and each of the words that make up the document is chosen from one of the topics. The topics are also probability distributed from words in a fixed dictionary.

The basic components of the LDA model are as follows [39]:

- Number of topics (K): This specifies the number of topics that the model will predict.
- Document topics (θ_d): Each document is represented as a mixture of K topics. Here, θ_d is the topic distribution of the d-th document.
- Topic word distributions (ϕ_k): Each topic represents a distribution of words. That is, ϕ_k is the word distribution of the k-th topic.
- Word vector (w_d, n): The words that make up each document. Each word is generated by a topic.

The mathematical structure of the LDA model can be expressed as follows:

$$p(w, z\ I\alpha, \beta) = \prod_{d=1}^{D}\left(\prod_{n=1}^{N_d} p\left(\left(w_{d,n}|z_{d,n}, \beta\right)p\left(z_{d,n}|\theta_d\right)\right)\right)p(\theta_d|\alpha)p(\beta|\eta) \quad (1)$$

where, w_d, n is the n-th word in the d-th document, and z_d, n is the predicted topic label for this word. θ_d represents the topic distribution of the d-th document. β and α are hyperparameters of Dirichlet distributions.

IV. RESULTS

Fig. 1 shows the block diagram of the study, illustrating the sequential stages of data collection, preprocessing, and analysis.



Fig. 1. Block diagram.

In total, 2046 tweets in Turkish and 18,000 tweets in English were initially collected from Twitter using relevant climate change-related keywords. The dataset underwent a rigorous preprocessing phase, including text cleaning, tokenization, and stop-word removal, to enhance the quality and reliability of the textual data. Following this process, tweets deemed unsuitable for analysis due to excessive noise, irrelevance, or duplication were excluded. As a result, the final dataset comprised 1104 Turkish tweets and 6449 English tweets that met the criteria for further sentiment and topic analysis.

The refined dataset was then subjected to deep learning-based sentiment classification and topic modelling to uncover prevalent themes and emotional tendencies in climate change discussions. The subsequent sections present the findings of the sentiment analysis, topic modelling results, and a comparative evaluation of Turkish and English tweets regarding climate change perceptions.

To classify the tweets according to their sentiment using machine learning methods, the data set was divided into two as training and test. Seventy percent of the data set is divided into training and 30 percent test data. At this stage, training and test sets were randomly generated to ensure data diversity. In all the algorithms used for sentiment analysis, 70% of the data set allocated for training is used while creating the model. After the model-building process, 30% of the test data is used to predict the model, and the performance of the model was evaluated.

A. Sentiment Analysis

Deep learning, a subset of machine learning, is used for sentiment analysis. Classification is the process of assigning a data set to the most appropriate one of the different and predetermined training data categories. The process of classifying the tweets according to their sentiments was carried out according to the climate change comments in the tweets. The aim of this study is to classify tweets about climate change with three emotions: positive, negative, and neutral.

Precision, recall, accuracy, and F1 metrics are used to evaluate the success of the model. TP (True Positives) refers to true positive predictions, FP (False Positives) refers to false positive predictions, FN (False Negatives) refers to false negative predictions, and TN (True Negatives) refers to true negative predictions [15]:

Precision indicates how accurate the result is in the model created.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall shows the ability to find the correct examples in the created model.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Accuracy indicates the success of the model in predicting the target classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

F1 Score is the harmonic mean of precision and recall.

$$F1 = \frac{2 * Recall * Precision}{Precision + Recall} \quad (5)$$

Table II shows the deep learning performance evaluation.

TABLE II. PERFORMANCE MEASUREMENT

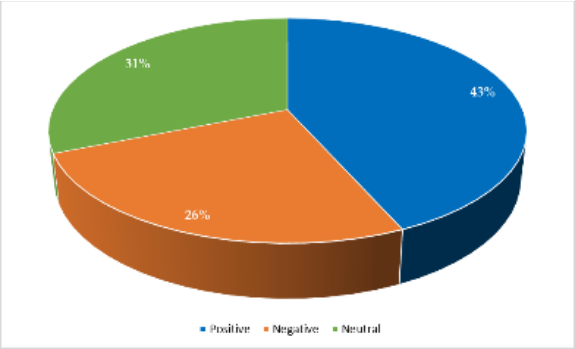
Metrics	English Tweet	Turkish Tweet
Precision	0.912	0.935
Recall	0.932	0.946
Accuracy	0.903	0.924
F Score (F1)	0.922	0.940

The deep learning model performed well on both English and Turkish tweets, with the metrics for Turkish tweets being generally better. Precision and recall values show that the model correctly captures positive classifications and misclassifications are low in both languages. For Turkish tweets, precision 0.935, recall 0.946 and F1 score 0.940 are higher than English tweets (0.912, 0.932 and 0.922, respectively). In terms of accuracy, Turkish tweets (92.4%) performed better than English tweets (90.3%). These results suggest that the language features or data structure of Turkish tweets may be better learnt by the model and that the model works more effectively with Turkish data. However, no statistical significance testing or confidence interval estimation was conducted. Future research should include hypothesis testing and variance analysis to assess the robustness of the observed performance differences. Table III shows the sentiment distribution of the tweets.

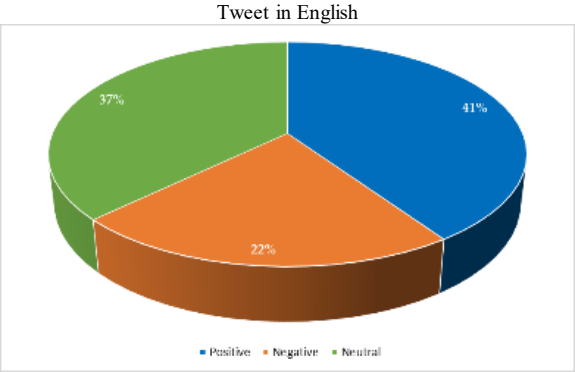
TABLE III. SENTIMENT STATISTICS

Sentiment Analysis	Tweet Number	
	EN	TR
Positive	2795	447
Negative	1646	246
Neutral	2007	411

The sentiment analysis results in Table III show that there are similar distributions between English and Turkish tweets. In both English and Turkish tweets, positive sentiment has a higher rate than other categories. Fig 2 shows the distribution of sentiment in terms of percentage.



Tweet in English



Tweet in Turkish

Fig. 2. Sentiment mood distribution.

In Fig. 2, 43.3% of the tweets in English were positive, 25.5% were negative, and 31.2% were neutral, while these rates were 40.5% positive, 22.3% negative, and 37.2% neutral in Turkish tweets, respectively. These distributions show that a generally positive or neutral perspective on climate change is dominant in both languages. However, the slightly higher rate of neutral sentiments in Turkish tweets compared to English tweets may suggest that Turkish-speaking users approach the issue with a more neutral or observant approach. In English tweets, the difference between positive and negative sentiments is more pronounced than in Turkish, which may indicate that English-speaking users may tend to have a stronger emotional response. However, this interpretation should be treated with caution, as the dataset is limited to Twitter users who shared posts using specific climate change-related keywords within a defined time frame.

B. Text Mining

The n-gram and term frequency and inverse term frequency methods are used to determine the core attributes. Word frequencies show the number of times words are used in the data set. Word frequencies are shown using a collection of matrix terms from the most frequently used words to the least frequently used words in the dataset. Term Frequency (TF) is the frequency of repetition of a term within a document. Since each document is different in length, a document is likely to appear much more often in longer documents than in shorter ones. Inverse Document Frequency (IDF) is the importance of a term in the entire document collection (D). According to IDF, the importance of a term is directly proportional to the frequency with which the term is used in the document and inversely proportional to the number of times the term is used in the entire document pool. Low-frequency terms have a high IDF score; high-frequency terms have a low IDF score. The term frequency-inverse text frequency (TF-IDF) value takes a high value if the term occurs a lot in a small number of documents. If the term is used in all documents, the TF-IDF value takes its lowest value [40].

Table IV shows the most frequent words extracted from English tweets related to the research topic.

TABLE IV. FREQUENCY LIST OF ENGLISH TWEETS

Positive	Negative	Neutral
(“China”, 153)	(“China”, 98)	(“China”, 185)
(“World”, 116)	(“Global”, 77)	(“Energy”, 152)
(“Energy”, 91)	(“Energy”, 49)	(“Global”, 50)
(“Carbon”, 77)	(“Money”, 39)	(“Money”, 48)
(“Natural”, 70)	(“Fossil”, 39)	(“Help”, 44)
(“Weather”, 61)	(“Government”, 37)	(“Warming”, 28)
(“Power”, 57)	(“Weather”, 36)	(“Weather”, 28)
(“Environmental”, 56)	(“Food”, 32)	(“Government”, 28)
(“Fossil”, 53)	(“Water”, 32)	(“Control”, 24)
(“Government”, 50)	(“Warming”, 32)	(“Carbon”, 22)

The analysis of positive, negative, and neutral words in English tweets in Table IV shows that discussions on climate change have a multidimensional structure. Among the positive words, terms such as “China” and “energy” point to promising developments such as renewable energy investments and carbon-neutral targets, while expressions such as “world” and “natural” reflect a positive view of global cooperation and environmentally friendly approaches. On the other hand, in the negative category, words such as “China,” “fossil,” and “government” represent criticism of carbon emissions, dependence on fossil fuels, and political inefficiencies. Neutral words, on the other hand, usually consist of terms used in neutral contexts such as “energy”, “global,” and “money,” and are used to describe news, policies, or general situation assessments related to climate change. These data reveal that, as stated in the literature [6], public perception of climate change is balanced between both positive and negative emotions and neutral information.

Table V shows the most frequent words extracted from Turkish tweets related to the research topic.

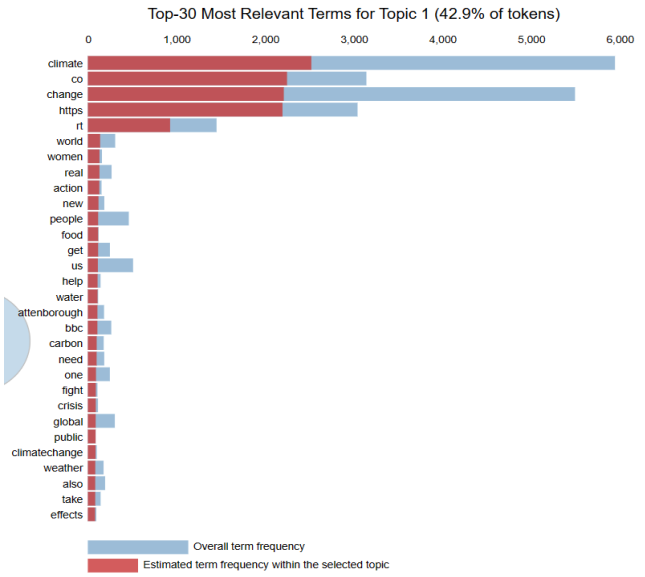
TABLE V. FREQUENCY LIST OF TURKISH TWEETS

Positive	Negative	Neutral
(“Environment,”, 106)	(“Global”, 60)	(“Urbanization”, 71)
(“Science”, 96)	(“Food”, 22)	(“Environment,”, 70)
(“Minister”, 89)	(“Companies”, 21)	(“Minister”, 49)
(“Urbanization”, 74)	(“Earthquake”, 17)	(“Presidency”, 32)
(“Global”, 40)	(“Environment,”, 15)	(“Education”, 18)
(“Warming”, 30)	(“Damage”, 13)	(“Disaster”, 31)
(“Water”, 24)	(“Warming”, 12)	(“Water”, 15)
(“Earth”, 24)	(“Water”, 12)	(“Management”, 12)
(“Economic”, 17)	(“Population”, 11)	(“Municipality”, 10)
(“Temperatures”, 14)	(“Carbon”, 9)	(“Cost”, 29)

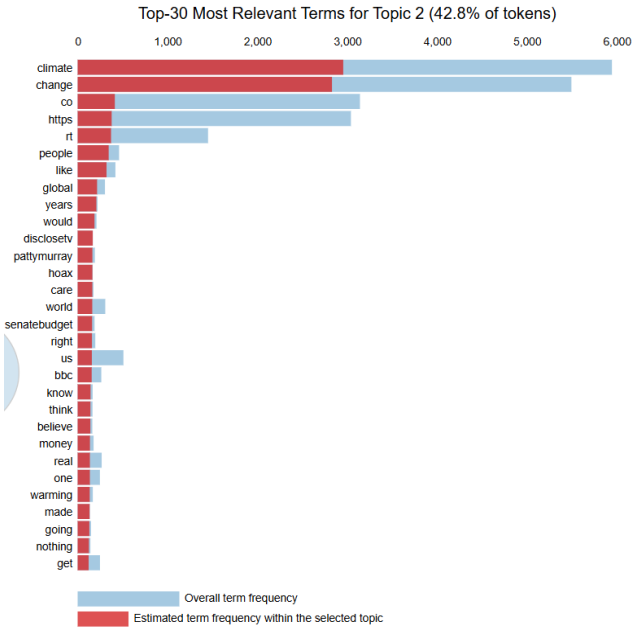
In Table V, the most common positive, negative, and neutral words in Turkish tweets in the context of climate change show that the perception of this issue has a multi-faceted structure. "Environment" and "science", which stand out among the positive words, reflect promising approaches to environmental and scientific issues, while terms such as "urbanization" and "minister" show that urbanization and initiatives taken at the leadership level are discussed in a positive framework. In contrast, words in the negative category, such as "global" and "food," represent concerns about global challenges and the food crisis. Expressions such as "earthquake" and "damage" indicate negative perceptions about the disasters triggered by climate change and their consequences. Neutral words such as "presidency," "management," and "disaster" are terms generally used in political, administrative, and crisis management discussions. The frequency of use of these words reveals that social and managerial sensitization about climate change has increased at both local and global levels. As stated in the literature [6], messages about climate change in the media and social media are mostly presented as a balance of both positive (solution-oriented approaches) and negative (worrisome effects) content. In this context, the discussions in Turkish tweets show that local problems (water resources, natural disasters) and solution proposals (scientific approaches, leadership) form a blended narrative.

C. Topic Modelling

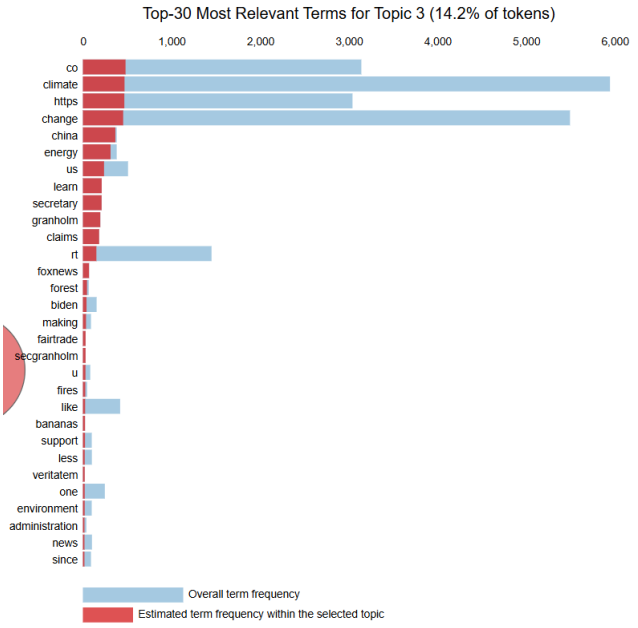
The algorithms proposed for topic modelling are statistical methods that aim to reach a conclusion by analyzing the words that make up the document. They do not need any labelling step while discovering the connection between the topics and the changes they show over time. While the placement of words in the document is ignored, the co-occurrence of words is used in this method. In this study, topic modelling is performed on 3 topics using Latent Dirichlet Allocation (LDA). The number of models was decided by trial and error, and the most appropriate number of topics was determined as 3. Fig. 3 shows the topic modelling of English tweets, and Fig. 4 shows the topic modelling of Turkish tweets.



Topic 1



Topic 2



Topic 3

Fig. 3. LDA terms (English tweets).

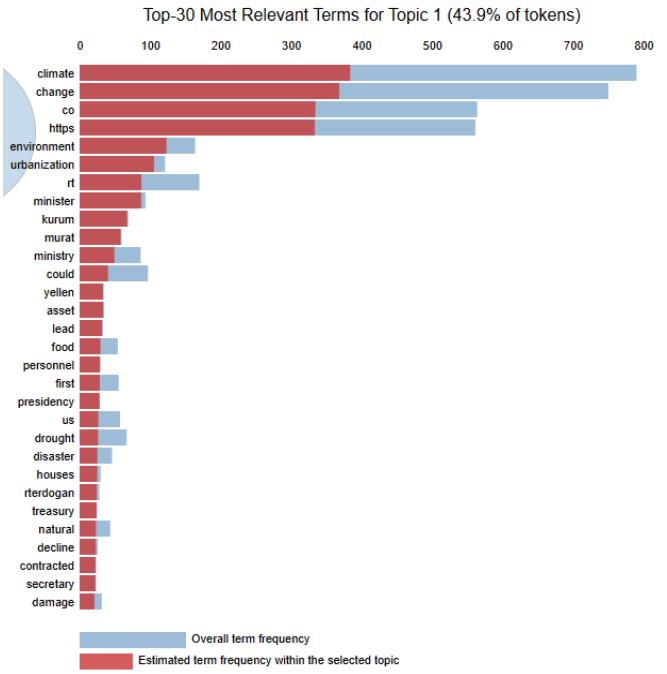
The three topic distributions of the English tweets in Fig. 3, which resulted from the topic modelling, are as follows:

Topic 1: Global Awareness and Calls to Action

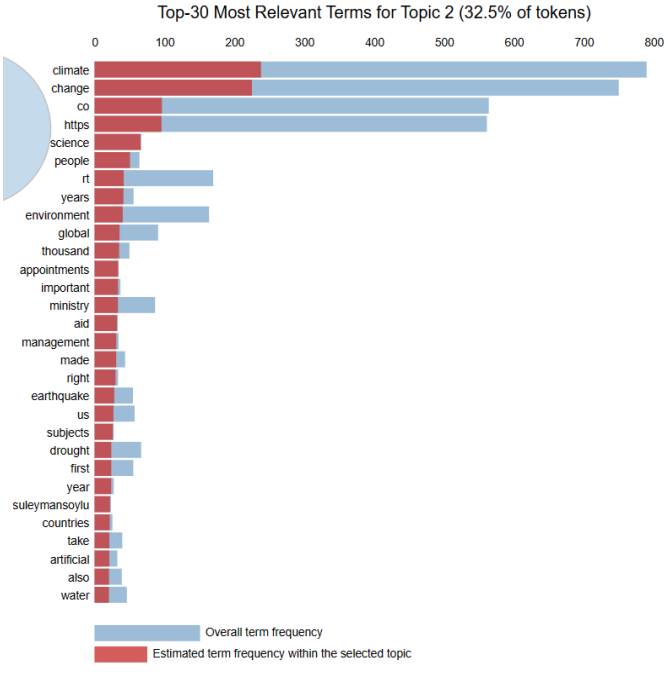
Topic 1 centers around awareness and calls for action on climate change. While the highest frequency of use of the term's "climate", and "change" constitutes the main theme of this topic, terms such as world, women, real, action, help, and public indicate that the global impacts of climate change are emphasized. Under this heading, awareness-raising and calls for action on the environmental crisis and broader societal issues come to the fore.

Topic 2: Climate Change Debates and Polarized Views

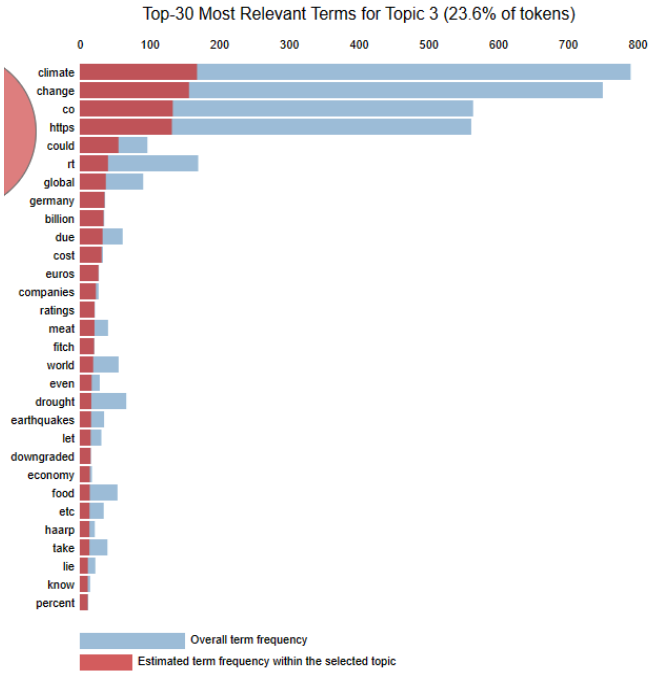
Topic 2 reflects the reality of climate change and the controversial discourses around it. While "climate", and "change" are again important keywords, terms such as "hoax", "believe", think", "care", "right", and "senate budget" points to the differences of opinion between those who believe in climate change and those who question it. In addition, terms such as money and economics indicate that the economic dimensions of climate policies also occupy an important place in the debates.



Topic 1



Topic 2



Topic 3

Fig. 4. LDA terms (Turkish tweets).

Topic 3: Political Discourses and Energy Policies

Topic 3 shows that climate change is addressed in the context of political actors, media, and energy policies. The prominence of terms such as "China", "energy", "secretary", "Granholm", "Fox News", and "administration" shows that climate change discourses are shaped through political figures, media organizations, and energy policies. Under this heading, discussions on energy policies and the role of specific countries or political actors come to the fore.

The three topic distributions of the Turkish tweets in Fig. 4, which resulted from the topic modelling, are as follows:

Topic 1: Political and Institutional Discourses and Climate Change

This topic reflects political, bureaucratic, and institutional discourses on climate change. Terms such as "minister", and "institution" indicate the intensification of public debates over the statements or policies of government officials and public institutions on this issue. Expressions such as "environment", "urbanization", and "natural disaster" reveal that environmental impacts, sustainable urbanization debates, and natural disasters are addressed in a political context.

Topic 2: Scientific Perspectives and Climate Crisis in a Global Framework

This heading includes scientific and global discourses on the climate crisis. Terms such as "science", "environment", and "global" indicate that scientific studies and international perspectives are frequently emphasized in tweets. In addition, terms such as "earthquake", "drought", and "water" emphasize natural disasters and resource problems caused by the climate crisis. In this context, the issue can be interpreted as an area

where climate change is discussed through scientific findings and global impacts.

Topic 3: Economic Impacts and Social Reactions

The third issue shows that climate change is related to economic impacts and societal perceptions. Expressions such as "economy", "cost", "companies", and "billion" reveal that economic risks and costs are frequently mentioned. Financial terms such as "downgraded" and "ratings" reflect the impacts of climate change on corporate and economic performance. In addition, terms such as "global" and "drought" indicate that this economic impact has broad repercussions at the global level.

V. Discussion

This study's analysis of social media perceptions of climate change makes important contributions to the literature by comparatively evaluating the contextual and emotional aspects of Turkish and English tweets. The findings obtained through deep learning-based sentiment analysis and topical modelling methods show that individuals" perceptions of climate change are influenced not only by linguistic but also cultural contexts. Accordingly, the study provides important results to better understand the nature of discussions on climate change in social media and to provide a roadmap for the development of environmental communication strategies.

The result of the study shows that emotional reactions were stronger in English tweets, while a more neutral attitude was more prominent in Turkish tweets. This finding coincides with the literature showing the relationship between cultural values, media usage habits and environmental sensitivities. [28] stated that social media plays an important role in public debates on climate change and that culture-specific perception differences are reflected in social media content.

While the higher proportion of neutral expressions in Turkish content indicates that users adopt an observant approach, the emotional intensity in English tweets reveals that the Western public adopts a more activist stance on environmental issues. This was also confirmed [37] in a cross-cultural study covering 110 countries; it was stated that the public perceived climate change as a higher risk and reactions to it were higher in more affected geographies.

Social media stands out as a space where individuals can make their voices heard on environmental issues, access information and play a role in shaping policies. In this context, the concept of "Fifth Power" articulated [10] emphasizes the capacity of individuals to generate knowledge and social intervention through digital platforms. In the tweets analyzed in this study, especially in English content, frequently used expressions (e.g., "action", "real", "help") can be considered as an indicator of digital activism.

In particular, the "Greta Effect", which symbolizes the impact of climate activist Greta Thunberg, has shown the power of online campaigns to mobilize decision-makers [20]. The findings obtained in the study reveal that similarly, large masses can be reached, and awareness can be raised through social media.

The use of machine learning and deep learning methods provides significant advantages in analyzing social media data. The LSTM model used in this study classified both Turkish and English tweets with high accuracy rates. This supports that deep learning architectures are more successful than traditional methods in text classification, as stated [35].

Higher F1 and accuracy rates for Turkish tweets indicate that the model may have learnt the language features better. This finding is in line with the studies conducted [3] on Turkish content classification. Although the Turkish dataset is smaller than the English dataset, the higher performance may be explained by differences in data characteristics. The Turkish dataset appears to be more homogeneous and less noisy, which can facilitate pattern learning for the model. In contrast, the larger English dataset likely includes more linguistic variability, sarcasm, and contextual ambiguity, which may reduce classification performance. Additionally, differences in class distribution may also contribute to this outcome.

The results of the topical modelling show that discussions on climate change are shaped around political, economic and scientific perspectives. While energy policies, media discourses and the role of political figures are prominent in English tweets, public institutions" statements and scientific discourses are more dominant in Turkish tweets. These differences reveal that environmental communication strategies should be customized according to language and context.

Especially words such as "hoax", "senate", and "budget" in English tweets reflect the polarization and political debates on climate change. [16] stated that such polarization weakens the consensus on climate change on social media and facilitates the spread of misinformation. These findings are limited to the analyzed dataset and should not be generalized to broader populations. The results reflect only the behavior of Twitter users within the selected time frame and keyword scope.

VI. Conclusions

The study provides an important perspective on this issue by analyzing social perceptions and emotional tendencies regarding climate change in Turkish and English tweets. Using artificial intelligence-based text mining and sentiment analysis methods, the study shows that discussions about climate change on social media platforms have a multidimensional structure. While the difference between positive (43.3%) and negative (25.5%) sentiments is more pronounced in English tweets, neutral sentiments (37.2%) have a higher rate in Turkish tweets. This suggests that English-speaking users tend to have stronger emotional reactions to climate change, while Turkish-speaking users tend to adopt a more observant or neutral approach. In the positive statements, words such as energy, China, and world point to promising developments such as renewable energy investments and global cooperation, while the words fossil, economy, and government in the negative category reflect criticisms such as fossil fuel dependency and political inefficiencies. These data support the findings in the literature that social perception of climate change is based on a balance of positive and negative emotions [37].

Nevertheless, these findings are specific to the sampled social media population and should not be generalized to broader linguistic communities. The results reflect only the behavior of users active on Twitter within the selected dataset and keyword scope.

In this study, deep learning-based models have been successfully used for sentiment analysis related to climate change. Natural language processing (NLP) techniques and deep learning methods were effectively applied to accurately classify the emotional content of Twitter data. The deep learning model LSTM (Long Short-Term Memory) offered higher accuracy rates than traditional methods in understanding the context of texts and accurately capturing emotional tones. In this way, it was possible to analyses users" feelings and thoughts about climate change in more depth and to compare social media data in different languages (Turkish and English). The high accuracy of deep learning methods in such analyses reveals the complex emotional structure of social media posts more clearly.

The importance of the article stems from the fact that it presents important implications at both the local and global levels by analyzing the perceptions and attitudes of individuals on climate change and climate change issues through social media data. Climate change is seen as a critical issue not only for environmental sustainability but also for the future of economic and social policies. Therefore, sentiment analysis of the content shared on social media about climate change provides valuable information to understand the public's concerns, hopes, and demands on this issue. The comparison of Turkish and English tweets has an important originality in terms of showing how perceptions towards climate change are shaped in different cultural contexts.

However, the study also has some limitations. Firstly, the fact that only Twitter data is used makes it difficult to generalize the results to other social media platforms. For example, a more comprehensive analysis can be made with content obtained from platforms such as Instagram, Facebook, or TikTok. Furthermore, the fact that only text-based analyses were conducted, and visual content or hashtag trends were not examined limited the possibility of making an evaluation covering all dimensions of social media posts. Another limitation is that the analysis is based on keyword frequencies regardless of the context of the texts. This may prevent a full understanding of users" real intentions and emotional states.

In future studies, comparative analyses of different social media platforms can be included, and the accuracy of sentiment analysis can be improved by using more advanced deep learning methods. In addition, the attitudes of different demographic groups (age, gender, geographical region) towards climate change can be analyzed, and more detailed inferences can be made on how local and global policies are perceived. By extending the study, more inclusive and effective strategies can be developed, especially for policymakers and stakeholders that drive climate change investments. Such analyses will play a critical role in both formulating policies in line with the demands of the public and raising awareness on combating climate change. Future studies may extend this work by incorporating multimodal data (e.g., images and videos) and developing real-time monitoring systems to track public perception of climate change.

REFERENCES

[1] IPCC, Climate Change 2023: Synthesis Report (Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change), 2023.

[2] S. Lewandowsky, J. Cook, N. Fay, and G. E. Gignac, "Science by social media: Attitudes towards climate change are mediated by perceived social consensus," Memory & Cognition, vol. 47, pp. 1445–1456, 2019.

[3] S. R. Weart, The Discovery of Global Warming, Harvard University Press, Cambridge, MA, 2005.

[4] S. C. Moser, "Communicating climate change: history, challenges, process and future directions," Wiley Interdisciplinary Reviews: Climate Change, vol. 1, no. 1, pp. 31–53, 2010.

[5] W. Pearce, S. Niederer, S. M. Özkula, and N. Sánchez Querubín, "The social media life of climate change: Platforms, publics, and future imaginaries," Wiley Interdisciplinary Reviews: Climate Change, vol. 10, no. 2, 2019.

[6] W. H. Dutton, The Fifth Estate: The Power Shift of the Digital Age, Oxford University Press, 2023.

[7] J. Bogert, J. Buczny, J. Harvey, and J. Ellers, "The effect of trust in science and media use on public belief in anthropogenic climate change: A meta-analysis," Environmental Communication, 2023.

[8] J. Thaker, "Cross-country analysis of climate change perception and protest participation," Environmental Communication, vol. 18, no. 3, pp. 233–249, 2024.

[9] N. G. Mede and R. Schroeder, "The "Greta effect" on social media: A systematic review," Environmental Communication, 2024.

[10] L. Dellmuth and K. Shyrokykh, "Climate change on Twitter: Implications for climate governance research," Wiley Interdisciplinary Reviews: Climate Change, vol. 14, no. 6, e848, 2023.

[11] J. Burger, M. Gochfeld, T. Pittfield, and C. Jeitner, "Perceptions of climate change and consequences," Energy and Power Engineering, vol. 8, no. 5, p. 250, 2016.

[12] L. Casiraghi, E. Kim, and N. Hara, "Tweeting on thin ice: Scientists in dialogic climate change communication," First Monday, 2024.

[13] J. Pueyo-Ros and E. Garau, "Characterizing attitudes towards climate change via sentiment analysis," Journal of Integrative Environmental Sciences, vol. 20, no. 1, 2023.

[14] L. Whitmarsh, "Risk perception and behavioural response to climate change," Journal of Risk Research, vol. 11, no. 3, pp. 351–374, 2008.

[15] A. Rousseau, "Climate-related social media engagement," Communication Research, 2023.

[16] X Platform, "X platform," 2024. [Online]. Available: https://x.com

[17] K. Makice, Twitter API: Up and Running, O"Reilly Media, 2009.

[18] M. Trupthi, S. Pabboju, and G. Narasimha, "Sentiment analysis on Twitter using streaming API," IEEE International Conference, 2017.

[19] M. Pejić Bach, Ž. Krstić, S. Seljan, and L. Turulja, "Text mining for big data analysis," Sustainability, vol. 11, no. 5, p. 1277, 2019.

[20] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing techniques for text mining," International Journal of Computer Science & Communication Networks, vol. 5, no. 1, pp. 7–16, 2015.

[21] M. Schonlau, N. Guenther, and I. Sucholutsky, "Text mining with n-gram variables," The Stata Journal, vol. 17, no. 4, pp. 866–881, 2017.

[22] B. Liu, Sentiment Analysis and Opinion Mining, Springer, 2022.

[23] E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, and A. R. Montejo-Ráez, "Sentiment analysis in Twitter," Natural Language Engineering, vol. 20, no. 1, pp. 1–28, 2014.

[24] Y. I. Alzoubi, A. E. Topcu, and A. E. Erkaya, "Machine learning-based text classification," Applied Sciences, vol. 13, no. 16, p. 9428, 2023.

[25] A. Palanivinayagam, C. Z. El-Bayeh, and R. Damaševičius, "Machine-learning-based text classification: A systematic review," Algorithms, vol. 16, no. 5, p. 236, 2023.

[26] I. H. Sarker, "Machine learning: Algorithms and applications," SN Computer Science, vol. 2, no. 3, p. 160, 2021.

[27] S. S. Nikam, "Classification techniques in data mining," Oriental Journal of Computer Science and Technology, vol. 8, no. 1, pp. 13–19, 2015.

[28] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification," in Data Classification: Algorithms and Applications, 2014.

[29] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models," Technical University of Denmark, 2012.

[30] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures," Artificial Intelligence Review, vol. 53, no. 6, pp. 4335–4385, 2020.

[31] S. F. Ahmed et al., "Deep learning modelling techniques," Artificial Intelligence Review, vol. 56, no. 11, pp. 13521–13617, 2023.

[32] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM hybrid model," IEEE Access, vol. 10, pp. 21517–21525, 2022.

[33] N. M. Rezk, M. Purnaprajna, T. Nordström, and Z. Ul-Abdin, "Recurrent neural networks," IEEE Access, vol. 8, pp. 57967–57996, 2020.

[34] J. Khan, N. Ahmad, S. Khalid, F. Ali, and Y. Lee, "Hybrid DNN for sentiment classification," IEEE Access, vol. 11, pp. 28162–28179, 2023.

[35] Y. Ho and S. Wookey, "Cross-entropy loss function," IEEE Access, vol. 8, pp. 4806–4813, 2019.

[36] V. Yesilevskyi and M. Kyt, "AI education trends," International Journal of Emerging Technologies in Learning, vol. 19, no. 4, pp. 86–96, 2024.

[37] K. Munir, H. Zhao, and Z. Li, "Semantic role labeling," IEEE/ACM Transactions, vol. 29, pp. 782–791, 2021.

[38] M. Mihalovic, "Classification models performance," Economics & Sociology, vol. 9, no. 4, p. 101, 2016.

[39] H. Jelodar et al., "Latent Dirichlet allocation and topic modeling," Multimedia Tools and Applications, vol. 78, pp. 15169–15211, 2019.

[40] H. Christian, M. P. Agus, and D. Suhartono, "TF-IDF for text mining," ComTech, vol. 7, no. 4, pp. 285–294, 2016.

Preference-Controllable Multi-Objective Deep Reinforcement Learning for Human-Robot Task Allocation in Service Environments

Asmaa Rashed Alahmari*, Wadee Alhalabi
Computer Science Department-Faculty of Computing & Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Human–Robot Collaboration (HRC) has gained increasing attention as it expands from industrial environments to service-oriented settings, where dynamic conditions and diverse operational objectives pose significant challenges for task allocation. Unlike controlled industrial environments, service contexts are characterized by frequent changes, uncertainty, and time-varying priorities, rendering static task allocation strategies ineffective. This paper proposes a method to address the problem of determining the optimal balance between human and robotic task allocation in dynamic service-oriented HRC systems. A preference-controllable multi-objective deep reinforcement learning framework is introduced to formulate task allocation as a dynamic, preference-dependent decision-making process. The proposed approach explicitly captures trade-offs among multiple, potentially conflicting objectives and enables adaptive task allocation under changing operational conditions and service priorities. The framework is evaluated through simulation-based experiments and comparative analysis with baseline strategies using multiple evaluation metrics, complemented by additional validation using external datasets. Experimental results demonstrate the effectiveness and adaptability of the proposed approach across varying preference configurations and workload conditions, supporting its applicability in real-world smart service environments.**

***Keywords*—Deep reinforcement learning; human–robot collaboration; preference-controllable reinforcement learning; smart service environments; task allocation**

I. Introduction

Human–Robot Collaboration (HRC) has demonstrated significant success in industrial environments, where humans and robots work collaboratively to enhance productivity, efficiency, and safety [1–3]. Building on this success, HRC has recently expanded beyond manufacturing into service-oriented sectors such as hospitality and tourism [4–12]. An increasing body of research has demonstrated the feasibility and benefits of deploying collaborative robots in service environments, reporting improvements in operational efficiency and service quality [4,6,11,12].

Among the core challenges in HRC systems, task allocation between human workers and robots has emerged as a critical research problem [13,14]. Determining which tasks should be performed by humans, robots, or jointly has been extensively investigated in industrial settings, both theoretically and experimentally [13–16].

Existing studies have proposed a wide range of task allocation approaches [13–21], including rule-based strategies, optimized heuristic methods, and more advanced solutions such as digital twins and machine learning–based models. More recently, deep reinforcement learning (DRL) has attracted growing attention as a promising approach for addressing complex and dynamic task allocation problems [15,22].

These approaches aim to optimize multiple objectives, such as efficiency, cost, performance, safety, and human workload [1,23–25]. While some studies focus on optimizing a single objective, others consider multiple objectives simultaneously [26–28]. Importantly, these objectives may be complementary in some cases, yet conflicting in others, leading to inherent trade-offs commonly described through Pareto optimality [29,30]. Moreover, task allocation strategies may be static or dynamic, depending on how objectives are defined and prioritized [31,32].

While task allocation in industrial environments benefits from relatively controlled and predictable settings, service environments present fundamentally different challenges [10]. Service systems are highly dynamic, characterized by frequent and unpredictable changes, diverse and often conflicting objectives, and time-varying priorities. For instance, speed and efficiency may be critical during peak hours, whereas customer experience and human touch become more important during off-peak periods. Unexpected events such as staff absence, equipment maintenance, or sudden demand fluctuations further complicate decision-making [5,8].

Despite the importance and complexity of task allocation in service-oriented HRC, existing research remains limited and lacks adaptive frameworks capable of dynamically balancing multiple conflicting objectives under changing preferences in real-world service settings [8,10].

To address these challenges, this paper proposes a preference-controllable multi-objective deep reinforcement learning approach for human–robot task allocation in service environments. Inspired by task allocation strategies developed for industrial HRC, the proposed framework enables dynamic adaptation to changing environmental conditions and shifting objective priorities. By explicitly modeling trade-offs among multiple objectives and allowing preference control, the framework aims to achieve a balanced allocation that improves efficiency, cost, and overall system performance in service-oriented contexts.

*Corresponding author.

The main contributions of this paper are as follows:

- Proposing a preference-controllable multi-objective deep reinforcement learning approach for dynamic human–robot task allocation in service environments.
- Modeling task allocation as a preference-dependent decision-making problem that captures trade-offs among conflicting objectives.
- Demonstrating adaptive and controllable allocation behavior under varying operational conditions and preference configurations.
- Validating the proposed approach using both simulation-based experiments and external real-world datasets.

This work contributes to determining the optimal balance between human and robotic task allocation in dynamic service-oriented HRC systems. Rather than assuming a fixed allocation, optimality is defined as a dynamic and preference-dependent trade-off among multiple objectives, enabling adaptive task allocation decisions under changing operational conditions and service priorities.

The remainder of this paper is organized as follows. Section II reviews the background and related work. Section III describes the proposed methodology. Section IV presents and discusses the experimental results. Section V outlines the limitations and directions for future research. Finally, Section VI concludes the paper.

II. Background and Related Work

This section presents the foundational concepts of the research and reviews existing studies on human–robot task allocation, multi-objective optimization, and preference-aware reinforcement learning, with a focus on identifying limitations in current approaches.

A. Task Allocation in Human–Robot Collaboration

Human–Robot Collaboration (HRC) refers to systems in which humans and robots work together within a shared workspace [13]. Research began in the early 1990s, initially focusing on human-compatible robotic hardware [18,33], and later expanded to include control modalities, interaction, and task representation [18], with comprehensive reviews provided in [34].

With Industry 4.0, HRC has rapidly grown in industrial applications, supported by safety standards such as ISO/TS 15066 and advances in machine learning for improved adaptability [18,35]. More recently, HRC has extended to service environments such as hospitality and retail [4–12], which are less structured, more dynamic, and strongly influenced by human behavior and demand, introducing additional challenges in coordination and decision-making [5,8].

Task allocation plays a central role in enabling effective HRC in both industrial and service environments, determining how responsibilities are shared between humans and robots while balancing objectives such as time, cost, safety, and workload [32]. Although widely studied in multi-robot systems [36], HRC introduces additional complexity due to human variability, making optimal allocation dynamic and context-dependent.

Early approaches focused on system-centric objectives such as makespan, cycle time, and cost [37–39], using methods including mixed-integer linear programming (MILP), meta-heuristics, and reinforcement learning [37–41]. More recent work has incorporated human-centered factors such as workload, ergonomics, and adaptability [42–47], with fatigue modeling approaches such as Learning-Forgetting-Fatigue-Recovery LFFR used to capture human performance dynamics [18,48–50].

Despite these advances, existing studies still lack worker personalization and balanced workload distribution, highlighting the importance of multi-objective optimization to balance human and system-level objectives [23,28]. Task allocation strategies can be categorized as static or dynamic [31,32]. Static approaches rely on predefined conditions and are sensitive to disruptions, while dynamic approaches adapt to real-time changes, making them more suitable for service environments such as cafés, where demand is variable and time-sensitive [32].

Among available techniques for enabling dynamic task allocation under uncertainty, reinforcement learning has emerged as a promising approach [32], including integration with digital twins and hybrid learning frameworks [20,51]. However, most existing work remains focused on industrial settings and does not address preference-controllable task allocation in dynamic service environments.

B. Multi-Objective Task Allocation and Pareto-Based Approaches

Multi-objective optimization is increasingly relevant in collaborative systems, as many real-world decision-making problems cannot be adequately represented by a single performance criterion. Prior work emphasizes that many real-world decision problems are characterized by multiple conflicting objectives that must be balanced based on their relative importance, and that most real-life problems are more naturally expressed with multiple objectives [29, 52]. In HRC, traditional approaches focus on system-centric metrics such as cycle time [28], which may negatively impact human factors such as fatigue and well-being.

Multi-objective task allocation addresses this by modeling trade-offs among competing objectives. Rather than a single optimal solution, a set of Pareto-optimal solutions is obtained, representing different trade-offs [28,29]. These solutions reflect varying priorities, where improving one objective may degrade another. Within HRC, Pareto-based approaches have been used to balance system efficiency and human factors, demonstrating that task allocation should be treated as a trade-off problem rather than a single-objective optimization [28].

Multi-objective reinforcement learning (MORL) extends this concept by learning policies that capture different trade-offs among objectives [29]. Approaches include single-policy methods, multi-policy methods, and preference-conditioned methods. While effective, these methods are often computationally expensive and not tailored to domain-specific

applications such as service-oriented HRC. Overall, although multi-objective approaches provide strong theoretical foundations, they are often abstract and lack direct application to dynamic, service-oriented human–robot task allocation problems.

C. Preference-Aware and Preference-Controllable Learning

In multi-objective problems, preferences are typically represented as weights indicating the importance of each objective [54]. Traditional approaches assume fixed weights, allowing reduction to single-objective formulations [52]. However, in practice, preferences are often unknown or change over time [30,52]. Preference-controllable reinforcement learning addresses this limitation by enabling policies to adapt to varying preference inputs [30]. Rather than learning separate policies, a single model is conditioned on preference vectors, allowing different trade-offs to be achieved dynamically [30,54].

In HRC systems, preferences are particularly important as they reflect human priorities and expectations. Prior studies have explored learning preferences through interaction and feedback [53], but these are generally not formulated as explicit task allocation problems and are not designed for service-oriented environments. Overall, existing approaches provide strong theoretical support for modeling dynamic preferences and adaptive behavior in multi-objective systems. However, these methods are not directly applied to task allocation in dynamic service-oriented HRC systems, where preferences may shift rapidly in response to workload, customer demand, or operational conditions.

D. Limitations of Existing Approaches

Despite progress, several limitations remain. First, most task allocation frameworks are designed for industrial environments and are not easily transferable to dynamic service settings [23,28]. Second, many approaches rely on fixed or implicit preference assumptions, limiting adaptability in real-world scenarios [52,54]. Third, there is a lack of unified frameworks that integrate task allocation, multi-objective optimization, and preference control into a single system.

Overall, existing studies address these aspects in isolation, such as industrial task scheduling, Pareto-based optimization, or preference-aware learning, but do not provide an integrated, preference-controllable framework for dynamic task allocation in service-oriented HRC systems, which motivates the proposed approach.

III. Methodology and Experimental Setup

This study proposes a formulation of human–robot task allocation as a preference-controllable multi-objective decision-making problem, where service priorities may vary dynamically across operational contexts. Rather than optimizing a single fixed objective or training multiple separate policies for different trade-offs, the proposed approach follows the principles of Preference-Controllable Reinforcement Learning (PreCo) [30]. In this framework, task allocation decisions are conditioned on a continuous preference vector that encodes the relative importance of operational efficiency, preservation of human-centric interactions, and effective resource utilization. By conditioning the policy on preferences, the agent can represent a continuum of task allocation strategies within a single model, enabling flexible adaptation to changing service priorities. Fig. 1 illustrates the overall preference-controllable learning framework adopted in this study, showing how user-defined trade-offs and estimated objective outcomes guide the agent's interaction with the environment.

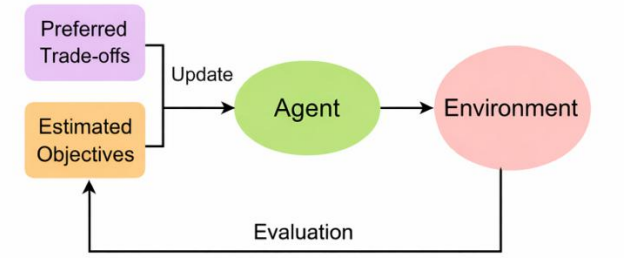


Fig. 1. Overview of the preference-controllable reinforcement learning (PreCo) framework adopted in this study.

A. Problem Formulation and Environmental Design

The task allocation problem is defined within a dynamic service-oriented HRC café environment, as a representative a real-world service scenario, where a human barista and a robotic agent jointly execute café service tasks under uncertain demand and variable execution conditions. At each decision step, the agent observes the current environment state and assigns the incoming order either to the human or to the robot, enabling adaptive allocation strategies as operational conditions evolve.

A custom simulation environment, CafeEnvironment, is developed using the Gymnasium interface [55]. Each episode represents a short operational window of 50 customer orders sampled from a predefined dataset comprising 20 heterogeneous order types (e.g., espresso-based drinks, specialty beverages, desserts, and packaged items). Each order is characterized by estimated human service time, robot service time, an automatability score $[0,1]$, and a human-touch importance score $[0,1]$. For food items, the estimated service times are modeled to be comprehensive, accounting for potential heating and custom preparation rather than mere handover. This task abstraction aligns with PreCo's vector-valued outcome formulation, where actions contribute differently to multiple objectives [30]. The dataset was constructed and validated through interviews with three café-domain experts (Table I).

B. Multi-Objective Learning Formulation

Task allocation performance is evaluated with respect to three potentially conflicting objectives: operational efficiency, preservation of human-centric interactions, and resource utilization, which represent core operational priorities in service-oriented café environments. These objectives are modeled using a vector-valued reward rather than a single scalar signal, explicitly capturing inherent trade-offs. Optimality is therefore defined relative to a preference vector specifying the relative importance of each objective, framing the problem as a preference-dependent multi-objective decision-making task. The policy is modeled as a conditional policy,

$$\pi(a \mid s, p),$$

TABLE I. ORDER TYPES AND TASK CHARACTERISTICS USED IN CAFEENVIRONMENT

Order Type	Human Service Time (s)	Robot Service Time (s)	Automatability	Human-Touch Importance
Espresso (single shot)	30	15	0.95	0.05
Double espresso / Lungo / Ristretto	30	16	0.92	0.08
Americano / Long Black	40	20	0.88	0.12
Cappuccino	60	35	0.80	0.20
Latte / Flat White / Macchiato / Cortado	60	36	0.78	0.22
Mocha / Hot Chocolate / Specialty hot drinks	90	45	0.72	0.30
Iced Coffee / Iced Latte / Iced Tea	90	50	0.70	0.28
Blended / Frappé beverages	90	55	0.60	0.35
Cold Brew / Nitro Cold Brew	30	30	0.40	0.70
Tea (bagged)	30	15	0.85	0.15
Tea (brewed, e.g., matcha latte)	60	30	0.68	0.32
Affogato / Dessert-coffee combinations	90	50	0.50	0.85
Muffin	180	120	0.30	0.55
Biscotti	180	120	0.30	0.45
Bagel	180	120	0.35	0.50
Croissant / Danish / Pastry	180	120	0.35	0.60
Cookie / Brownie / Cake slice	180	120	0.30	0.45
Sandwich	180	120	0.35	0.55
Wrap / Salad bowl	180	120	0.35	0.50
Yogurt parfait / Fruit cup	180	120	0.35	0.50

where s denotes the environment state, a the allocation action, and p the preference vector. Conditioning the policy on preferences follows the preference-conditioned policy formulation adopted in [30] and allows a single model to represent a continuum of allocation strategies without retraining.

C. State, Action, and Preference Modeling

At each decision step t, the action space is discrete with two possible actions:

$$a_t \in \{0,1\},$$

where $a_t = 0$ assigns the incoming order to a human barista and $a_t = 1$ assigns it to the robotic system.

The environment state is represented as a 10-dimensional continuous vector normalized to $[0,1]$, combining operational context, task characteristics, system utilization, and preference information:

$$s_t = [w_t, o_t, h_t, m_t, u_H, u_R, p_{\text{eff}}, p_{\text{touch}}, p_{\text{util}}, c],$$

Here, w_t denotes the remaining workload in the episode expressed as the fraction of unprocessed orders; o_t represents the normalized categorical index of the current order type; h_t and m_t correspond to the human-touch importance and automatability scores of the task, respectively. The variables u_H and u_R denote the current utilization ratios of the human worker and the robotic system. The preference vector $p = (p_{\text{eff}}, p_{\text{touch}}, p_{\text{util}})$ encodes the relative importance assigned to operational efficiency, preservation of human-centric interactions, and resource utilization. Finally, c represents the operational load context, discretized into low, medium, and high demand levels. Including preferences directly in the observation space follows the core principle of preference-conditioned reinforcement learning adopted in PreCo [30], enabling the policy to adapt its behavior to different trade-off configurations without modifying the underlying environment dynamics. At the beginning of each episode, a preference vector is sampled and kept fixed throughout the episode, allowing the policy to learn consistent allocation behavior under a specific preference setting.

D. Reward Design and Multi-Objective Structure

At each decision step t, the environment provides a vector-valued reward that captures multiple aspects of task allocation performance, including service efficiency, preservation of human-centric interactions, and resource utilization balance. Representing rewards in vector form allows each objective to be modeled and analyzed independently before aggregation. To ensure transparency, the individual reward components of operational efficiency (r_t^{eff}), human-touch preservation (r_t^{touch}), and resource utilization (r_t^{util}) are formally defined based on the environment dynamics. The efficiency reward is defined as:

$$r_t^{eff} = \gamma_c \left(1 - \min\left(\frac{T(a_t)}{T_{\max}}, 1\right)\right)$$

where $T(a_t)$ denotes the service time of the selected agent (human or robot), $T_{\max}$ is the maximum service time observed in the system, and γ_c is a context-dependent scaling factor reflecting workload conditions (γ_c= 1.2 for high load, 1.0 for medium load, and 0.9 for low load).

The human-touch reward captures the importance of human interaction (h_t) and is defined conditionally based on the selected agent a_t. When a human is assigned ($a_t = 0$):

$$r_t^{touch} = \min\left(h_t \cdot (1 + 0.5\, p_{touch}) + \delta, 1\right)$$

where $\delta = 0.12\, p_{touch}$ if $p_{touch} > 0.6$, and $\delta = 0$ otherwise. When a robot is assigned ($a_t = 1$):

$$r_t^{touch} = 1 - \min\left(\text{penalty}, 1\right)$$

where the penalty term is defined as: penalty = $0.7\, h_t\, (p_{touch} - 0.6) \times 4.5$, if $p_{touch} > 0.6$, penalty = $0.2\, h_t p_{touch}$, otherwise.

The utilization reward promotes balanced workload distribution between human and robot agents and is defined as:

$$r_t^{util} = 1 - |\, u_H - u_R\, |$$

where u_H and u_R denote the current utilization ratios of the human and robot, respectively.

Finally, let $p = (p_{\text{eff}}, p_{\text{touch}}, p_{\text{util}})$ represent the sampled preference vector specifying the relative importance of each objective. The overall scalar reward used for policy optimization is computed through preference-weighted linear scalarization as follows:

$$R_t = p_{\text{eff}}\, r_t^{\text{eff}} + p_{\text{touch}}\, r_t^{\text{touch}} + \beta_{\text{util}}\, p_{\text{util}}\, r_t^{\text{util}},$$

where β_{util} is a scaling parameter controlling the contribution of the utilization term relative to the other objectives.

Preference-awareness is further incorporated during training through a lightweight shaping mechanism applied outside the environment. Following the Preference-Controllable Reinforcement Learning (PreCo) paradigm [30], an additional preference-alignment signal based on changes in cosine similarity between the cumulative objective vector and the sampled preference vector is used to guide learning. This shaping mechanism is implemented at the wrapper level and does not modify the PPO loss function or the underlying network architecture.

E. Learning Algorithm and Evaluation

The agent is trained using Proximal Policy Optimization (PPO) as implemented in Stable-Baselines3 [56], chosen for its stability and robustness in stochastic environments. Observation and reward normalization are applied using VecNormalize. Key hyperparameters include a learning rate of 3×10^{-4}, discount factor $\gamma = 0.99$, GAE parameter $\lambda = 0.95$, clip range 0.2, and batch size 128. These hyperparameters were selected based on established best practices for PPO and widely adopted configurations in prior reinforcement learning studies, and were empirically validated to ensure stable convergence in the proposed task allocation environment. Training is conducted for 300,000 interaction steps, corresponding to approximately 6,000 episodes.

Following common practice in preference-controllable reinforcement learning, evaluation focuses on two main requirements [30]. The first is the ability of the agent to explore the Pareto front and represent diverse trade-offs among objectives. The second is controllability, defined as the ability of the learned policy to produce a performance trade-off that aligns with the input preference. These requirements are evaluated using Hypervolume (HV) to assess Pareto front coverage and a similarity measure $\Psi(p, v_\pi)$ to quantify preference alignment. In addition to these preference-based metrics, task-level performance is assessed using episodic measures such as average task completion time, makespan, and human–robot allocation ratios, which are commonly used evaluation metrics in human–robot collaborative task allocation studies. All experiments are conducted with fixed random seeds to ensure reproducibility.

IV. Results and Discussion

This section presents and analyzes the experimental results obtained from both internal simulation-based evaluations and external validation experiments, along with their implications for the proposed framework.

A. Internal Evaluation Results

The results in this section are obtained after training the proposed preference-controllable policy, demonstrating stable and consistent learning behavior over 6,000 training episodes. To evaluate the controllability of the learned policy beyond training-time preference sampling, we conduct a controlled evaluation under three representative preference scenarios, designed to reflect realistic operational conditions in a service environment. These scenarios correspond to three dominant service modes commonly encountered in practice.

The first scenario is efficiency-oriented, representing high-demand periods (peak-hour scenarios) where rapid service and throughput are prioritized, and robotic execution is preferred for handling the majority of tasks. The second scenario is human-touch-oriented, emphasizing customer interaction and human touch, where task allocation favors human baristas while still allowing a balanced distribution with robotic assistance when appropriate. The third scenario is balanced, targeting a balanced utilization of human and robotic resources to ensure efficient operation without excessive workload concentration or prolonged idleness.

This evaluation protocol mirrors the controlled preference experiments commonly adopted in preference-controllable reinforcement learning studies, where fixed and interpretable preference vectors are used to examine policy behavior under specific trade-off configurations, as discussed in [30]. To analyze how the learned agent adapts its decisions under each scenario, we report the following evaluation metrics: preference alignment (Ψ), hypervolume (HV), and human–robot task allocation behavior.

In the following subsections, we analyze each of these aspects in detail to demonstrate the controllability, interpretability, and robustness of the proposed framework.

1) Preference alignment (Ψ): Preference alignment (Ψ) is used to evaluate how closely the learned policy follows the intended preference direction under fixed and interpretable preference configurations. It is computed as the cosine similarity between the cumulative achieved objective vector and the corresponding preference vector, providing a directional measure of controllability consistent with preference-controllable reinforcement learning.

As shown in Fig. 2, the balanced preference configuration exhibits the strongest alignment ($\Psi = 0.983$), indicating that the policy effectively tracks the intended trade-off when multiple objectives are jointly emphasized. In contrast, lower alignment values are observed under efficiency-oriented ($\Psi = 0.513$) and human-touch-oriented ($\Psi = 0.675$) preferences. These configurations correspond to more asymmetric preference settings, where emphasis is placed predominantly on a single objective.

This behavior is expected and reflects a fundamental characteristic of preference-controllable reinforcement learning as discussed in [30]. During training, preferences are sampled continuously across the preference simplex, encouraging the policy to remain responsive across a broad range of trade-offs rather than specializing toward extreme or corner configurations. Consequently, alignment under highly asymmetric preferences is naturally lower than under balanced configurations, without indicating a loss of controllability.

Overall, the variation in preference alignment across scenarios demonstrates that the learned policy responds systematically and directionally to changes in the preference input. Rather than converging to a fixed objective or ignoring preference information, the policy adapts its behavior in accordance with the specified preference configuration, satisfying a core requirement of controllable multi-objective decision-making in human–robot collaboration settings.

2) Pareto Quality (Hypervolume, HV): To complement the preference alignment analysis, we evaluate the Pareto solution quality of the learned policy using the hypervolume (HV) metric, which measures the volume of the objective space dominated by the achieved solutions. Hypervolume is a standard indicator in multi-objective optimization, capturing both convergence and coverage of the Pareto front.

Fig.3 reports the HV values obtained under the three controlled preference scenarios. The balanced configuration achieves the highest hypervolume ($HV \approx 0.41$), followed by the efficiency-oriented ($HV \approx 0.31$) and human-touch-oriented ($HV \approx 0.26$) scenarios. This pattern indicates that balanced preferences enable solutions that preserve more favorable trade-offs across objectives, resulting in broader Pareto coverage.

In contrast, preference configurations that strongly emphasize a single objective naturally constrain the solution space, leading to reduced hypervolume. This behavior is consistent with multi-objective optimization theory, where prioritizing one objective limits the achievable trade-offs among the remaining objectives. Importantly, lower HV values under efficiency- or human-touch-oriented preferences do not indicate suboptimal learning, but rather reflect the intentional focus imposed by the specified preference direction. These results align with observations reported in preference-controllable reinforcement learning literature, where balanced or moderately weighted preferences tend to yield higher Pareto quality, while extreme preferences trade overall solution diversity for targeted objective satisfaction.

3) Human–robot task allocation behavior: To examine how preference alignment and Pareto quality translate into operational behavior, we analyze the resulting human–robot task allocation under the controlled preference scenarios. Fig. 4 illustrates the proportion of tasks assigned to the robot and the human agent for efficiency-oriented, human-touch-oriented, and balanced preferences.

Under the efficiency-oriented configuration, the policy allocates a substantially larger share of tasks to the robot (approximately 0.85 robot utilization), reflecting an automation-driven strategy that prioritizes service speed and throughput during peak operational periods. This behavior is consistent with the specified preference direction and confirms that the policy leverages robotic execution when efficiency is emphasized.

In contrast, the human-touch-oriented configuration results in a pronounced shift toward human execution, with robot utilization decreasing to approximately 0.36. Tasks associated with higher interaction importance are preferentially handled by human baristas, demonstrating that the policy internalizes task semantics and adjusts allocation decisions to preserve human-centric service quality when required.

The balanced configuration converges to a moderate allocation (approximately 0.54 robot utilization), representing a stable compromise between automation and human involvement. This allocation pattern aligns with the higher preference alignment and hypervolume observed under balanced preferences, indicating that joint emphasis on multiple objectives yields both robust Pareto performance and interpretable allocation behavior.

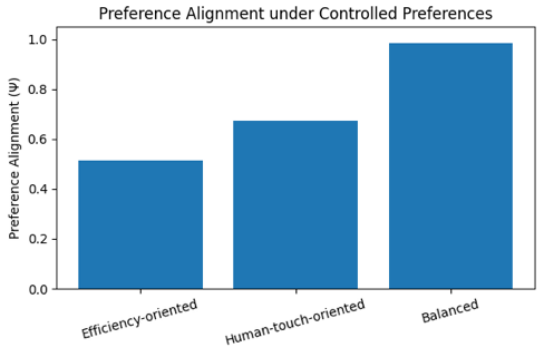


Fig. 2. Preference alignment (Ψ) under controlled preference configurations.

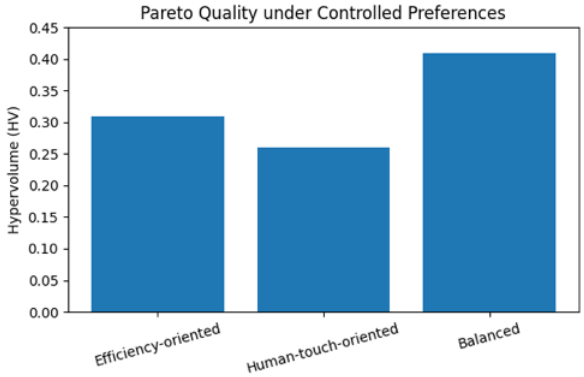


Fig. 3. Pareto quality measured by hypervolume (HV) under controlled preference configurations.

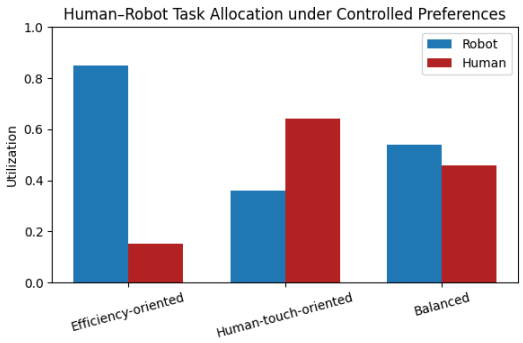


Fig. 4. Human–robot task allocation behavior under controlled preference configurations.

Notably, robot participation remains non-negligible across all scenarios. This outcome reflects the structure of the task environment, in which a subset of orders exhibits high automatability and favorable robot execution times. Rather than suppressing automation under human-centric preferences, the proposed framework permits robot involvement whenever it remains consistent with task characteristics and the specified preference constraints. Consequently, observed changes in utilization reflect relative behavioral adaptation rather than absolute enforcement of automation or manual operation.

Overall, these results demonstrate that the learned policy does not converge to a fixed automation ratio nor rely on static heuristics. Instead, it dynamically adjusts human–robot roles in a directionally consistent and interpretable manner in response to changing preferences, satisfying a central requirement of controllable human–robot collaboration in service-oriented environments.

The observed evaluation patterns are consistent with prior findings in preference-controllable reinforcement learning, where both preference alignment (Ψ) and Pareto quality (HV) tend to decrease under extreme or single-objective-dominant preference configurations, while more balanced preferences yield higher alignment and more favorable multi-objective trade-offs. In our experiments, this behavior manifests clearly as increased robot utilization under efficiency-oriented preferences, reduced robot involvement under human-touch-oriented preferences, and higher Ψ and HV under balanced configurations, demonstrating that the proposed framework exhibits controllable, preference-aware behavior in a service-oriented human–robot collaboration environment with interpretable task allocation decisions.

4) Comparison with baseline strategies: To assess the effectiveness of the proposed framework, we conduct a comparative evaluation against several baseline strategies. The comparison focuses on efficiency-oriented performance, measured by mean makespan and average task execution time, as well as the resulting human–robot allocation ratios.

Specifically, the proposed PreCo+PPO approach is evaluated under the three representative preference configurations defined in the previous section, which are efficiency-oriented, human-touch-oriented, and balanced, to reflect different operational priorities in service environments. These configurations are compared against four baseline strategies commonly used in practice: a rule-based two-threshold heuristic, full human execution, full robot execution, and a static 50/50 allocation. The rule-based two-threshold strategy assigns tasks using fixed decision rules based on task characteristics, favoring robot execution for highly automatable tasks and human execution for tasks requiring higher human involvement, without adapting to changing workloads or preferences. The full human and full robot strategies assign all tasks exclusively to a single agent, representing fully manual and fully automated operations, respectively. The static 50/50 strategy maintains an equal task split between human and robot throughout the episode, without considering task characteristics or operational context.

This comparison aims to examine whether the proposed approach can achieve competitive efficiency relative to automation-heavy strategies, while simultaneously maintaining meaningful human involvement and adaptive task allocation behavior that other baseline methods cannot provide. The results in Fig. 5 and Table II highlight the advantages of the proposed preference-controllable approach over baseline strategies. Under the efficiency-oriented configuration, the PreCo+PPO policy achieves a substantially lower mean makespan compared to rule-based and static allocation strategies, approaching the performance of the full-robot baseline while avoiding the complete elimination of human involvement. This demonstrates that high operational efficiency can be achieved without resorting to extreme automation, which has been associated with known limitations in prior studies.

When prioritizing human touch, the framework intentionally increases reliance on human execution to support customer interaction and service quality. As expected, this shift leads to a higher makespan due to reduced automation. However, compared to the full-human baseline, which results in the longest makespan and average task time, the proposed approach still benefits from selective robotic assistance to reduce overall service time while maintaining a strong level of human participation. This indicates that the framework does not simply replace automation with manual execution, but instead balances human involvement with robotic support to mitigate known challenges of fully manual operation, such as prolonged processing times, workload variability, and sensitivity to human-related disruptions reported in prior studies.

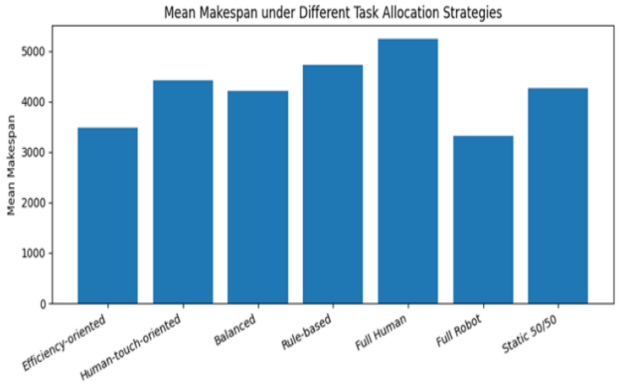


Fig. 5. Mean makespan comparison across preference-controllable strategies and baseline methods. Lower values indicate higher efficiency.

TABLE II. COMPARISON OF PREFERENCE-CONTROLLABLE AND BASELINE TASK ALLOCATION STRATEGIES

Category	Strategy	Mean Makespan	Avg Task Time (s)	Robot %	Human %
PreCo+PPO	Efficiency-oriented	3476.439	69.529	0.856	0.144
	Human-touch-oriented	4414.963	88.299	0.366	0.634
	Balanced	4207.118	84.142	0.547	0.453
Baseline	Rule-based	4730.312	94.606	0.417	0.583
	Full Human	5240.619	104.812	0.000	1.000
	Full Robot	3320.709	66.414	1.000	0.000
	Static 50/50	4268.116	85.362	0.510	0.490

The balanced configuration provides a practical trade-off between efficiency and human involvement. It achieves lower makespan values than the rule-based and static baselines while maintaining a near-equal human–robot allocation ratio. The slightly lower makespan compared to the static 50/50 strategy suggests that adaptive preference-aware control can offer modest efficiency gains over fixed allocation ratios under similar human–robot balance targets.

Overall, these results demonstrate that the proposed preference-controllable framework enables adaptive and interpretable task allocation across diverse operational priorities. Unlike static or heuristic-based methods, the proposed approach allows explicit control over trade-offs between efficiency and human-centric considerations without retraining or redesigning the policy, highlighting its practical applicability in real-world service-oriented environments.

B. External Validation Results

After establishing internal validity through controlled simulated experiments, we further evaluate the external validity of the proposed framework using real-world transactional data. The objective of this evaluation is to assess the robustness of the learned policy when exposed to realistic, heterogeneous, and noisy demand distributions that differ substantially from the synthetic order streams used during training. Importantly, the external evaluation does not modify the environment dynamics, reward formulation, or learning architecture. The trained PreCo+PPO policy is kept fixed throughout all experiments, and only the incoming order sequences are replaced with those observed in real transaction logs. This design allows the effect of realistic demand patterns on policy behavior to be isolated without introducing confounding factors related to retraining or architectural changes.

To assess external validity across different operational conditions, three real-world coffee sales datasets are employed [57-59], as summarized in Table III. These datasets differ in scale, structure, and data source, ranging from a large-scale transactional dataset comprising 149,116 transactions to smaller datasets with a few thousand records each. All datasets contain complete transactional information and exhibit demand characteristics commonly observed in real coffee shop operations.

Exploratory analysis shows that customer demand follows a skewed, long-tailed distribution, with coffee and tea products accounting for most orders, followed by bakery and specialty items. Most transactions include one or two items, indicating that the datasets largely reflect individual service requests rather than bulk orders. These characteristics are consistent with the task-level abstraction used in the simulated café environment. To integrate the real-world data, each transaction is mapped to the predefined service task categories using a simple rule-based mapping that preserves the original order sequence and task characteristics.

External evaluation follows the same controlled protocol used during internal validation. Three fixed and interpretable preference configurations are evaluated under deterministic execution, and performance is assessed using preference alignment (Ψ), Pareto hypervolume (HV), and human–robot utilization ratios. A quantitative summary of preference-aligned performance across both internal and external datasets is reported in Table IV, while Fig. 6 visualizes the results across datasets.

Across all external datasets and preference configurations, the proposed framework exhibits consistent, robust, and interpretable behavior. Under efficiency-oriented preferences, the policy consistently favors robotic execution, achieving robot utilization rates of approximately 83–85% across all datasets. Preference alignment and Pareto quality remain comparable to, and in several cases higher than, internal validation results, indicating effective prioritization of service efficiency under realistic and noisy demand conditions. In contrast, human-touch-oriented preferences result in a pronounced shift toward human execution, with robot utilization decreasing to approximately 35–38% across all datasets. This behavior remains stable across datasets of different sizes and sources, confirming that the policy responds to preference information rather than overfitting to specific order distributions.

The balanced preference configuration consistently yields the highest preference alignment ($\Psi \approx 0.97$–0.99) and competitive hypervolume values across all datasets, producing stable and interpretable human–robot workload distributions. Minor variations between internal and external evaluations are expected given the skewed and noisy nature of real-world transactional data and do not indicate performance degradation.

Overall, the external validation results demonstrate that the learned policy does not overfit to synthetic task distributions and maintains stable, interpretable, and preference-aligned behavior when deployed under realistic service demand compositions. The consistency of behavioral trends across datasets of different scales and sources provides strong evidence of external validity and supports the applicability of the proposed preference-

controllable framework to real-world human–robot collaborative service environments.

C. Discussion

Building on the experimental results, this section discusses the implications of preference-controllable reinforcement learning for human–robot task allocation in service environments. The findings demonstrate that a single preference-conditioned policy can adapt its allocation behavior in response to varying operational priorities while maintaining stable learning dynamics and interpretable decision patterns.

A key observation is that balanced preference configurations consistently yield stronger preference alignment (Ψ) and higher Pareto quality (HV) compared to extreme or single-objective-dominant scenarios. This behavior aligns with established findings in preference-controllable reinforcement learning, where policies trained to respond across a distribution of preferences tend to achieve more stable and well-balanced trade-offs under moderate settings. In contrast, extreme preferences naturally constrain the solution space, reducing Pareto coverage and reflecting an inherent trade-off between specialization and controllability rather than a limitation of the proposed approach.

Importantly, these behavioral patterns are preserved under external validation using real-world transactional datasets. Despite differences in scale, structure, and demand variability between synthetic and real-world data, the learned policy exhibits consistent directional responses to preference changes. This indicates that the policy captures task allocation principles that remain stable under realistic distributional shifts, rather than relying on environment-specific patterns.

Beyond aggregate performance metrics, the results provide insight into how preference-controllable learning translates into interpretable operational behavior. Efficiency-oriented preferences lead to increased robot utilization and faster task execution, while human-touch-oriented preferences shift allocation toward human workers to support service interaction and personalization. Balanced preferences produce moderate robot utilization, reflecting a practical compromise between automation and human involvement. These systematic allocation patterns are observed consistently across internal and external evaluations, confirming that the policy responds coherently to preference inputs rather than relying on static heuristics or fixed allocation rules.

Notably, robot participation remains non-negligible across all preference scenarios. This outcome is driven by the task composition of service environments, where many tasks exhibit high automatability and favorable robot execution characteristics. Rather than suppressing automation under human-centric preferences, the proposed framework allows robotic participation whenever it aligns with task requirements and preference constraints. This design reflects realistic service operations, in which automation and human labor coexist dynamically rather than being treated as mutually exclusive.

From both a practical and conceptual perspective, these findings suggest that preference-controllable reinforcement learning provides a flexible mechanism for managing human–robot collaboration in service settings where objectives may shift over time or differ across stakeholders. Adjusting preference weights enables controlled changes in operational behavior without retraining or redesigning the policy, supporting adaptability to changing service priorities and contextual requirements.

Overall, the discussion highlights that the primary contribution of this work lies not in optimizing a single performance metric, but in demonstrating how preference-controllable reinforcement learning can support controllable, interpretable, and context-aware task allocation in human–robot collaborative services. By explicitly modeling human touch alongside efficiency and utilization, and validating behavior under realistic demand conditions, this work bridges multi-objective reinforcement learning with practical service-oriented human–robot collaboration, highlighting its applicability in real-world deployment scenarios.

TABLE III. SUMMARY OF DATASETS USED FOR INTERNAL AND EXTERNAL VALIDATION

Dataset	Size (Transactions)	Reference
Dataset 1	149,116	[57]
Dataset 2	3,636	[58]
Dataset 3	3,898	[59]

TABLE IV. SUMMARY OF PREFERENCE-ALIGNED PERFORMANCE ACROSS INTERNAL AND EXTERNAL DATASETS

Dataset	Scenario	Preference Alignment (Ψ)	Hypervolume (HV)	Robot Utilization
Simulated Dataset	Efficiency-oriented	0.513	0.305	0.85
	Human-touch-oriented	0.675	0.264	0.36
	Balanced	0.983	0.408	0.54
Dataset 1	Efficiency-oriented	0.597	0.426	0.84
	Human-touch-oriented	0.653	0.290	0.37
	Balanced	0.969	0.485	0.56
Dataset 2	Efficiency-oriented	0.521	0.355	0.84
	Human-touch-oriented	0.730	0.196	0.37
	Balanced	0.987	0.349	0.58
Dataset 3	Efficiency-oriented	0.515	0.357	0.83
	Human-touch-oriented	0.732	0.196	0.37
	Balanced	0.987	0.347	0.58

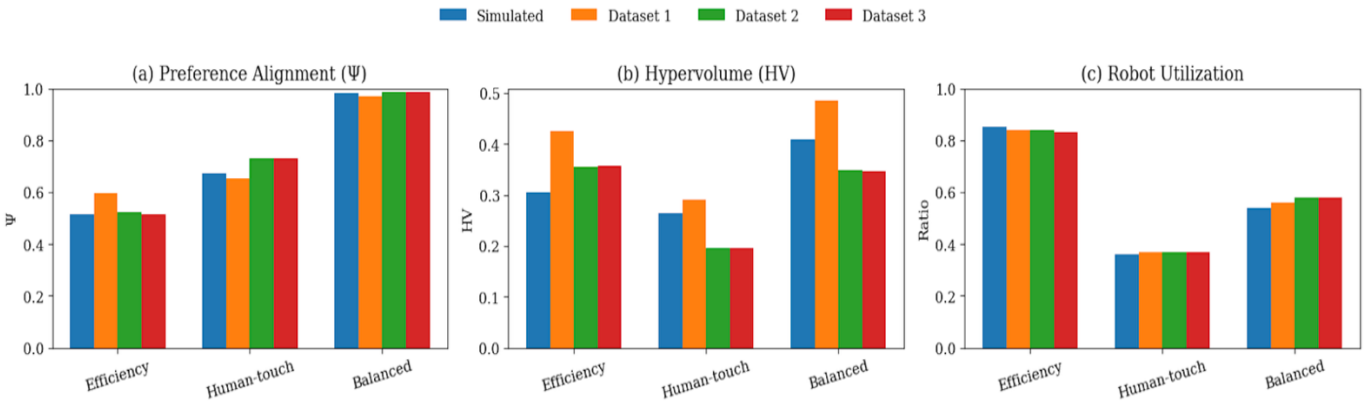


Fig. 6. Preference-controllable performance across datasets.

V. Limitations and Future Research

While the proposed framework demonstrates effective preference-controllable task allocation in a service-oriented environment, several limitations should be acknowledged. First, the task environment considered in this study contains a relatively high proportion of automatable tasks with favorable robot execution times. Although this reflects many contemporary service settings, environments with a larger share of human-exclusive or socially sensitive tasks may exhibit different allocation dynamics. Future work could explore alternative task compositions and service domains to assess the robustness of the proposed approach across a broader range of human-centered service contexts.

Second, the current study assumes explicitly specified and fixed preference configurations during evaluation, enabling controlled analysis of preference-dependent behavior. These configurations correspond to three common and practically relevant service scenarios in café environments, namely efficiency-oriented operation during peak hours, human-touch-oriented service during customer-focused or off-peak periods, and a balanced mode representing routine daily operation. In real service environments, however, preference priorities may shift gradually in response to operational conditions, customer feedback, or managerial decisions. Extending the framework to support smoothly varying or dynamically inferred preferences, while preserving policy stability and interpretability, represents an important direction

Third, although the simulation environment is designed to capture realistic service characteristics and demand patterns, the evaluation remains simulation-based. Real-world deployment may introduce additional sources of uncertainty, including human behavioral variability, execution delays, and operational constraints that are not explicitly modeled. In addition, task attributes such as automatability and human-touch importance are assumed to be known *a priori*. In this study, these attributes are derived from expert-informed estimates based on interviews with café domain specialists; however, in practice, such characteristics may be uncertain, noisy, or context-dependent. The simulated environment further assumes equal availability of human and robotic resources, adopting a 1:1 human-to-robot ratio to enable controlled analysis of task allocation behavior. Accordingly, future work will explore alternative human–robot resource compositions beyond the assumed 1:1 ratio, including configurations such as 1:2 or 2:1, to examine how varying workforce balance influences task allocation behavior and system performance.

Beyond these extensions, future research will focus on deploying the proposed framework in real or hybrid service settings involving human operators, learning task attributes online, and incorporating adaptive preference updates driven by human or customer feedback. Further extensions to more complex service scenarios, such as multiple human workers, heterogeneous robotic agents, or dynamically evolving task sets, would also enhance the applicability and scalability of preference-controllable human–robot collaboration beyond simulated environments.

VI. Conclusion

This paper introduced a preference-controllable reinforcement learning framework for human–robot task allocation in service-oriented environments. By integrating multi-objective reward design with preference-conditioned policy learning, the proposed framework enables a single trained policy to adapt task allocation behavior across varying efficiency, human-touch, and resource utilization priorities without requiring retraining. The experimental results demonstrate stable learning behavior, clear preference-dependent allocation patterns, and interpretable shifts between human and robotic execution under different operational priorities. Rather than enforcing fixed automation ratios, the framework supports adaptive task allocation that reflects both task characteristics and user intent, allowing human–robot collaboration strategies to be adjusted in response to changing service demands.

Validation across both simulated environments and real-world transactional datasets shows that the learned policy maintains consistent and interpretable behavior under realistic and heterogeneous demand conditions. These findings indicate that preference-controllable reinforcement learning provides a practical and flexible mechanism for managing trade-offs between efficiency and human-centric considerations in service settings. Overall, this work bridges multi-objective and preference-aware reinforcement learning with practical human–robot collaboration, demonstrating how preference-driven control can support adaptive, transparent, and context-aware

task allocation in real-world service environments, highlighting its potential for real-world deployment.

ACKNOWLEDGMENT

The project was funded by KAU Endowment (WAQF) at king Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for technical and financial support.

REFERENCES

[1] M. Dhanda, B. A. Rogers, S. Hall, E. Dekoninck, and V. Dhokia, "Reviewing human-robot collaboration in manufacturing: Opportunities and challenges in the context of Industry 5.0," Robotics and Computer-Integrated Manufacturing, vol. 93, p. 102937, 2025.

[2] A. Baratta, A. Cimino, M. G. Gnoni, and F. Longo, "Human–robot collaboration in Industry 4.0: A literature review," Procedia Computer Science, vol. 217, pp. 1887–1895, 2023.

[3] S. K. Yadav and S. Shahi, "Safe human–robot collaboration in dynamic environments: An AI-powered situation awareness perspective," International Journal of Tropical Medicine, vol. 19, pp. 92–98, 2024.

[4] Y. Choi, M. Choi, M. Oh, and S. Kim, "Service robots in hotels: Understanding service quality perceptions of human–robot interaction," Journal of Hospitality Marketing & Management, vol. 29, no. 6, pp. 613–635, 2020.

[5] G. G. Liu, P. Benckendorff, and G. Walters, "Human–robot interaction research in hospitality and tourism: Trends and future directions," Tourism Review, 2024.

[6] T. Shimmura, R. Ichikari, T. Okuma, H. Ito, K. Okada, and T. Nonaka, "Service robot introduction to a restaurant enhances both labor productivity and service quality," Procedia CIRP, vol. 88, pp. 589–594, 2020.

[7] W. Grobbelaar, A. Verma, and V. K. Shukla, "Analyzing human–robot interaction in the food industry," Journal of Physics: Conference Series, vol. 1714, p. 012032, 2021.

[8] R. de Kervenoael, R. Hasan, A. Schwob, and E. Goh, "Leveraging human–robot interaction in hospitality services: Incorporating the role of perceived value, empathy, and information sharing into visitors' intentions to use social robots," Tourism Management, vol. 78, p. 104042, 2020.

[9] H. Qiu, M. Li, B. Shu, and B. Bai, "Enhancing hospitality experience with service robots: The mediating role of rapport building," Journal of Hospitality Marketing & Management, vol. 29, no. 3, pp. 247–268, 2020.

[10] I. Tussyadiah, "A review of research into automation in tourism: Launching the Annals of Tourism Research curated collection on artificial intelligence and robotics in tourism," Annals of Tourism Research, vol. 81, p. 102883, 2020.

[11] H.-W. Jang and S.-B. Lee, "Serving robots: Management and applications for restaurant business sustainability," Sustainability, vol. 12, no. 10, p. 3998, 2020.

[12] M. Xie and H.-B. Kim, "User acceptance of hotel service robots using the quantitative Kano model," Sustainability, vol. 14, no. 7, p. 3988, 2022.

[13] Y. Y. Liau and K. Ryu, "Task allocation in human–robot collaboration based on task characteristics and agent capability for mold assembly," Procedia Manufacturing, vol. 51, pp. 179–186, 2020.

[14] R. Mammadzada, "Various scheduling techniques used in human–robot collaborative system strategies for allocation and optimization of tasks," Journal of Modern Technology & Engineering, vol. 8, no. 3, 2023.

[15] J. Ding, M. Chen, T. Wang, J. Zhou, X. Fu, and K. Li, "A survey of AI-enabled dynamic manufacturing scheduling: From directed heuristics to autonomous learning," ACM Computing Surveys, vol. 55, no. 14s, pp. 1–36, 2023.

[16] M. L. Lee, X. Liang, B. Hu, G. Onel, S. Behdad, and M. Zheng, "A review of prospects and opportunities in disassembly with human–robot collaboration," Journal of Manufacturing Science and Engineering, vol. 146, no. 2, p. 020902, 2024.

[17] E. Merlo, E. Lamon, F. Fusaro, M. Lorenzini, A. Carfì, F. Mastrogiovanni, and A. Ajoudani, "An ergonomic role allocation framework for dynamic human–robot collaborative tasks," Journal of Manufacturing Systems, vol. 67, pp. 111–121, 2023.

[18] G. Bruno and D. Antonelli, "Dynamic task classification and assignment for the management of human–robot collaborative teams in workcells," The International Journal of Advanced Manufacturing Technology, vol. 98, no. 9, pp. 2415–2427, 2018.

[19] B. Tian, M. Janardhanan, and M. Marinelli, "A systematic investigation of the barriers to effective implementation of human–robot assembly lines: An integrated multi-criteria decision-making approach," International Journal of Computer Integrated Manufacturing, vol. 37, no. 1–2, pp. 198–223, 2024.

[20] W. Ren, X. Yang, Y. Yan, Y. Hu, and L. Zhang, "A digital twin-based framework for task planning and robot programming in human–robot collaboration," Procedia CIRP, vol. 104, pp. 370–375, 2021.

[21] Q. Lv, R. Zhang, X. Sun, Y. Lu, and J. Bao, "A digital twin-driven human–robot collaborative assembly approach in the wake of COVID-19," Journal of Manufacturing Systems, vol. 60, pp. 837–851, 2021.

[22] W. Hou, Z. Xiong, M. Yue, and H. Chen, "Human–robot collaborative assembly task planning for mobile cobots based on deep reinforcement learning," Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, vol. 238, no. 23, pp. 11097–11114, 2024.

[23] C. Urrea, "Hybrid deep learning–reinforcement learning for adaptive human–robot task allocation in Industry 5.0," Systems, vol. 13, no. 8, p. 631, 2025.

[24] C. Petzoldt, D. Niermann, E. Maack, M. Sontopski, B. Vur, and M. Freitag, "Implementation and evaluation of dynamic task allocation for human–robot collaboration in assembly," Applied Sciences, vol. 12, no. 24, p. 12645, 2022.

[25] A. Ali, H. Azevedo-Sa, D. M. Tilbury, and L. P. Robert Jr., "Heterogeneous human–robot task allocation based on artificial trust," Scientific Reports, vol. 12, no. 1, p. 15304, 2022.

[26] M. Faccio, I. Granata, and R. Minto, "Task allocation model for human–robot collaboration with variable cobot speed," Journal of Intelligent Manufacturing, vol. 35, no. 2, pp. 793–806, 2024.

[27] R. Wang, D. Zhao, A. Gupte, and B. C. Min, "Initial task allocation in multi-human multi-robot teams: An attention-enhanced hierarchical reinforcement learning approach," IEEE Robotics and Automation Letters, vol. 9, no. 4, pp. 3451–3458, 2024.

[28] S. Chand and Y. Lu, "Dual task scheduling strategy for personalized multi-objective optimization of cycle time and fatigue in human–robot collaboration," Manufacturing Letters, vol. 35, pp. 88–95, 2023.

[29] J. Xu, Y. Tian, P. Ma, D. Rus, S. Sueda, and W. Matusik, "Prediction-guided multi-objective reinforcement learning for continuous robot control," in Proc. International Conference on Machine Learning (ICML), PMLR, Nov. 2020, pp. 10607–10616.

[30] Y. Yang, T. Zhou, M. Pechenizkiy, and M. Fang, "Preference controllable reinforcement learning with advanced multi-objective optimization," in Proc. 42nd International Conference on Machine Learning (ICML), 2025.

[31] A. Alessio, K. Aliev, and D. Antonelli, "Multicriteria task classification in human–robot collaborative assembly through fuzzy inference," Journal of Intelligent Manufacturing, vol. 35, no. 5, pp. 1909–1927, 2024.

[32] C. Shyalika, T. Silva, and A. Karunananda, "Reinforcement learning in dynamic task scheduling: A review," SN Computer Science, vol. 1, no. 6, p. 306, 2020.

[33] E. Helms, R. D. Schraft, and M. Hägele, "rob@work: Robot assistant in industrial environments," in Proc. IEEE Int. Workshop on Robot and Human Interactive Communication, 2002, pp. 399–404.

[34] M. A. Goodrich and A. C. Schultz, "Human–robot interaction: A survey," Foundations and Trends in Human–Computer Interaction, vol. 1, pp. 203–275, 2007.

[35] D. Nguyen-Tuong and J. Peters, "Model learning for robot control: A survey," Cognitive Processing, vol. 12, no. 4, pp. 319–340, 2011.

[36] A. Khamis, A. Hussein, and A. Elmogy, "Multi-robot task allocation: A review of the state of the art," pp. 31–51, Springer International Publishing, 2015.

[37] C. Weckenborg, K. Kieckhäfer, C. Müller, M. Grunewald, and T. S. Spengler, "Balancing of assembly lines with collaborative robots," Business Research, vol. 13, no. 1, pp. 93–132, 2019, doi: 10.1007/s40685-019-0101-y.

[38] C. Ferreira, G. Figueira, and P. Amorim, "Scheduling human–robot teams in collaborative working cells," International Journal of Production Economics, vol. 235, p. 108094, 2021, doi: 10.1016/j.ijpe.2021.108094.

[39] F. Chen, K. Sekiyama, F. Cannella, and T. Fukuda, "Optimal subtask allocation for human and robot collaboration within hybrid assembly systems," IEEE Transactions on Automation Science and Engineering, vol. 11, pp. 1065–1075, 2014, doi: 10.1109/TASE.2013.2274099.

[40] A. Raatz, S. Blankemeyer, T. Recker, D. Pischke, and P. Nyhuis, "Task scheduling method for human–robot collaboration workplaces based on capabilities and execution time assumptions for robots," CIRP Annals, vol. 69, pp. 13–16, 2020, doi: 10.1016/j.cirp.2020.04.030.

[41] T. Yu, J. Huang, and Q. Chang, "Optimizing task scheduling in human–robot collaboration with deep multi-agent reinforcement learning," Journal of Manufacturing Systems, vol. 60, pp. 487–499, 2021, doi: 10.1016/j.jmsy.2021.07.015.

[42] M. Breque, L. De Nul, and A. Petridis, Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry, 2021.

[43] Y. Wang, H. S. Ma, J. H. Yang, and K. S. Wang, "Industry 4.0: A way from mass customization to mass personalization production," Advances in Manufacturing, vol. 5, pp. 311–320, 2017, doi: 10.1007/s40436-017-0204-7.

[44] Y. Lu, H. Zheng, S. Chand, W. Xia, Z. Liu, X. Xu, et al., "Outlook on human-centric manufacturing towards Industry 5.0," Journal of Manufacturing Systems, vol. 62, pp. 612–627, 2022, doi: 10.1016/j.jmsy.2022.02.001.

[45] E. Matheson, R. Minto, E. G. G. Zampieri, M. Faccio, and G. Rosati, "Human–robot collaboration in manufacturing applications: A review," Robotics, vol. 8, p. 100, 2019, doi: 10.3390/robotics8040100.

[46] A. Cherubini, R. Passama, A. Crosnier, A. Lasnier, and P. Fraisse, "Collaborative manufacturing with physical human–robot interaction," Robotics and Computer-Integrated Manufacturing, vol. 40, pp. 1–13, 2016, doi: 10.1016/j.rcim.2015.12.007.

[47] L. Wang, R. Gao, J. Váncza, J. Krüger, X. V. Wang, S. Makris, et al., "Symbiotic human–robot collaborative assembly," CIRP Annals, vol. 68, pp. 701–726, 2019, doi: 10.1016/j.cirp.2019.05.002.

[48] M. Pearce, B. Mutlu, J. Shah, and R. Radwin, "Optimizing makespan and ergonomics in integrating collaborative robots into manufacturing processes," IEEE Transactions on Automation Science and Engineering, vol. 15, pp. 1772–1784, 2018, doi: 10.1109/TASE.2018.2789820.

[49] M. Y. Jaber, Z. S. Givi, and W. P. Neumann, "Incorporating human fatigue and recovery into the learning–forgetting process," Applied Mathematical Modelling, vol. 37, pp. 7287–7299, 2013, doi: 10.1016/j.apm.2013.02.028.

[50] K. Li, Q. Liu, W. Xu, J. Liu, Z. Zhou, and H. Feng, "Sequence planning considering human fatigue for human–robot collaboration in disassembly," Procedia CIRP, vol. 83, pp. 95–104, 2019.

[51] J. Wang, Y. Yan, Y. Hu, X. Yang, and L. Zhang, "A transfer reinforcement learning and digital-twin based task allocation method for human–robot collaboration assembly," Engineering Applications of Artificial Intelligence, vol. 144, p. 110064, 2025.

[52] A. Abels, D. Roijers, T. Lenaerts, A. Nowé, and D. Steckelmacher, "Dynamic weights in multi-objective deep reinforcement learning," in Proc. International Conference on Machine Learning (ICML), PMLR, May 2019, pp. 11–20.

[53] M. D. Zhao, R. Simmons, and H. Admoni, "Learning human contribution preferences in collaborative human–robot tasks," in Proc. Conference on Robot Learning (CoRL), PMLR, Dec. 2023, pp. 3597–3618.

[54] F. Felten, L. N. Alegre, A. Nowé, A. Bazzan, E. G. Talbi, G. Danoy, and B. C. da Silva, "A toolkit for reliable benchmarking and research in multi-objective reinforcement learning," Advances in Neural Information Processing Systems, vol. 36, pp. 23671–23700, 2023.

[55] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, et al., "Gymnasium: A Standard Interface for Reinforcement Learning Environments," in The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2026. [Online]. Available: https://openreview.net/forum?id=qPMLvJxtPK. Accessed: Jan. 2026.

[56] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," Journal of Machine Learning Research, vol. 22, no. 268, pp. 1–8, 2021.

[57] A. Abbas, "Coffee Sales Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/ahmedabbas757/coffee-sales. Accessed: Jan. 2026.

[58] R. Richard, "Coffee Store Sales," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/reignrichard/coffee-store-sales. Accessed: Jan. 2026.

[59] I. Helon, "Coffee Sales Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/ihelon/coffee-sales. Accessed: Jan. 2026.

Fuzzy-Integrated Modular Neural Networks for Accurate Prediction of On-Time-In-Full Supply Chain Performance

Mariem Mrad[1], Mohamed Amine Frikha[2*], Younes Boujelbene[3], Soufiene Ben Othman[4]
Applied College, King Faisal University, 31982 Al-Ahsa, Saudi Arabia,[2,4]
Faculty of Economics and Management of Sfax, Tunisia[1,3]

***Abstract*—Accurate prediction of supply chain performance is essential for improving operational efficiency and enabling proactive, data-driven decision-making under dynamic and uncertain conditions. Conventional forecasting methods often struggle to capture the nonlinear relationships between operational factors and performance outcomes. This paper proposes an improved neural modeling framework for predicting supply chain performance based on artificial neural networks (ANNs). The proposed approach compares a mono-network model (global ANN) with a modular multi-network architecture composed of several local neural models integrated through a fuzzy fusion mechanism. Unlike existing studies that focus on isolated performance metrics, this work targets the prediction of the key composite indicator On-Time-In-Full (OTIF). Simulation experiments conducted on a nonlinear dynamic supply chain system demonstrate that the modular ANN approach achieves a significant reduction in learning error, dropping from 0.0223 in the global model to as low as 0.0004 in local modules. Furthermore, the total training time was reduced from 1631.58 seconds to an average of approximately 311 seconds per module. These results confirm that fuzzy-integrated modular architectures offer superior generalization and computational efficiency for advanced predictive analytics in complex supply chain management (SCM) environments.**

***Keywords*—Artificial neural networks; supply chain management; performance prediction; predictive analytics; modular-network modeling; demand forecasting**

I. Introduction

Reliable prediction of supply chain performance has emerged as a key requirement for organizations operating in volatile, complex, and highly interconnected markets, where competitiveness and sustainability depend on proactive decision-making [1]. As noted by Chopra and Meindl [4], as supply chains grow more digitized and globally distributed, their performance exhibits strong non-linearities, multi-criteria interdependencies, and time-varying dynamics. These are characteristics that traditional predictive approaches often cannot capture effectively. Classical techniques such as time-series modeling, econometric regression, and analytical optimization offer important insights but generally assume linear structures, static relationships, or single-output formulations, limiting their applicability to modern, data-rich supply chain environments.

Advances in artificial intelligence, particularly neural network–based methods, have significantly transformed predictive analytics in SCM [2]. Neural networks are capable of approximating highly nonlinear functions and capturing complex temporal dependencies within multidimensional data streams. More recently, research has evolved beyond conventional multilayer perceptron's toward hybrid deep learning architectures, such as CNN-LSTM, GNN-LSTM, and attention-based models, that integrate spatial, temporal, and relational information [3]. These approaches leverage the rich data ecosystems enabled by modern smart manufacturing technologies, such as interconnected sensors, intelligent automation systems, and digital twin frameworks, thereby supporting adaptive, real-time, and data-driven decision-making processes.

The concept of "supply chain performance" has evolved beyond traditional operational and financial metrics. Contemporary frameworks increasingly integrate Environmental, Social, and Governance (ESG) dimensions, promoting sustainable and socially responsible operations. As a result, composite metrics combining service-level indicators such as On-Time-In-Full (OTIF), cost efficiency, and inventory measures with sustainability criteria have emerged as holistic representations of supply chain outcomes [4].

Despite significant progress, several gaps persist. Most models focus on single variables, neglecting the interdependencies among performance dimensions. Furthermore, existing studies commonly employ a single global neural network, which struggles to model distinct operating regimes, such as stable periods versus disruption episodes [5]. Consequently, this study addresses the following central research question: How can a fuzzy-integrated modular neural architecture enhance the predictive accuracy, generalization capability, and computational efficiency of multi-dimensional supply chain KPIs compared to a traditional monolithic neural network model?

In response to these issues, this paper proposes an improved neural modeling framework for simultaneously predicting a vector of core supply chain KPIs, namely OTIF, total cost, and inventory level [6]. The contribution is twofold. First, a modular multi-network architecture is developed, in which specialized neural models learn the behavior of different operational zones. Second, these local models are integrated using a fuzzy fusion mechanism that dynamically weights predictions according to the prevailing system state [7]. This design mirrors the inherent regime-dependent behavior of real supply chains and enables a

*Corresponding author.

more robust approximation of their underlying nonlinear dynamics.

This research aims to:

- Develop artificial neural network models to capture composite supply chain performance through nonlinear system approximation.
- Develop and implement a modular multi-network architecture based on specialized local models that capture zone-specific dynamics.
- Integrate local models using a fuzzy fusion mechanism that adapts to operational conditions.
- Perform a thorough empirical evaluation comparing the single-network (global) model and the proposed multi-network architecture in terms of accuracy, generalization, and training efficiency.

The results demonstrate that the multi-network architecture significantly outperforms the traditional mono-network model, offering a more robust and accurate approach for capturing the dynamic, nonlinear, and multidimensional aspects of supply chain performance. This framework contributes both methodologically and practically to advancing next-generation predictive analytics in complex supply chain environments.

The remainder of this paper is structured as follows. Section II reviews related work on neural network–based prediction methods and composite supply chain performance modeling. Section III presents the system modeling framework and formulates the performance prediction problem. Section IV describes the proposed neural modeling approach, including the single-network and multi-network architectures and the fuzzy fusion mechanism. Section V outlines the experimental setup, simulation results, and comparative analysis of the proposed models. Finally, Section VI concludes the paper and discusses potential directions for future research.

II. Related Work

Neural network applications in SCM have undergone a significant paradigm shift, transitioning from basic feedforward structures to modular and hybrid deep learning systems. This evolution is driven by the need to handle the increasing volume, velocity, and variety of data generated in modern logistics networks [7-8].

A. Evolution Neural Architecture: fFrom MLP to Hybrid Models

Early research in supply chain predictive analytics relied heavily on the Multilayer Perceptron (MLP) and standard Backpropagation algorithms [9]. While these models proved effective for simple function approximation, they often failed to capture the long-term temporal dependencies and spatial relationships inherent in global supply networks. To overcome these limitations, recent literature has pivoted toward hybrid architectures [10]. For instance, CNN-LSTM models have emerged as a robust solution, utilizing Convolutional Neural Networks (CNNs) to extract local spatial features while Long Short-Term Memory (LSTM) units handle temporal sequences.

Expanding beyond temporal data, Graph Neural Networks (GNNs) integrated with LSTMs (GNN-LSTM) have gained prominence for their ability to represent supply chains as complex graphs. These models allow for the modeling of multi-hop dependencies and inter-organizational relationships, providing a more accurate reflection of how disruptions propagate through a network [11-12]. Furthermore, the integration of attention mechanisms and transformer-based models has further enhanced the ability of neural systems to focus on critical operational drivers during periods of high volatility.

B. Evolution of Composite Supply Chain Performance Metrics

The conceptualization of supply chain performance has evolved from isolated operational indicators toward multidimensional composite performance frameworks. Traditional evaluation models, such as the Supply Chain Operations Reference (SCOR) model and the Balanced Scorecard (BSC), emphasize integration across operational efficiency, financial outcomes, and customer service dimensions [13]. These frameworks provide structured methodologies for measuring service reliability, responsiveness, cost control, and asset utilization.

More recently, Environmental, Social, and Governance (ESG) criteria have been incorporated into supply chain performance assessment. ESG indicators increasingly serve as leading predictors of supply chain resilience, risk exposure, and long-term sustainability [14]. This shift reflects growing recognition that operational performance cannot be evaluated independently of environmental impact, regulatory compliance, and stakeholder expectations.

However, despite the emergence of composite evaluation frameworks, most neural network–based predictive models continue to focus on single performance indicators. The modeling of interdependent KPIs, such as OTIF, total cost, and inventory levels, remains relatively underexplored. This gap highlights the need for multi-output and modular architectures capable of capturing interrelated performance dynamics across heterogeneous operational regimes.

C. Training Optimization and Feature Engineering

The effectiveness of any neural predictive model is deeply tied to its training optimization and the quality of input features. While the Levenberg–Marquardt algorithm is widely recognized for its superior convergence speed in nonlinear system identification, it remains susceptible to local optima in highly complex supply chain datasets. Consequently, hybrid optimization techniques, such as combining gradient descent with metaheuristics like Particle Swarm Optimization (PSO) or Genetic Algorithms (GA), have become essential for hyperparameter tuning [15].

Moreover, advanced feature engineering, including the use of autoencoders for dimensionality reduction and automated feature selection, has improved model expressiveness. These strategies allow models to filter noise in high-dimensional data environments, ensuring that the neural network learns the underlying operational dynamics rather than over-fitting to transient fluctuations in demand or lead times.

D. Integration with Industry 4.0 Technologies

The proliferation of Industry 4.0 technologies has transformed the data ecosystem in which neural models operate. The integration of Internet of Things (IoT) sensors, RFID tracking, and blockchain-enabled traceability provides a continuous stream of real-time data [16]. One of the most critical advancements in this domain is the emergence of the Digital Twin (DT).

Digital twins provide virtual replicas of physical supply chain processes, allowing neural models to perform "what-if" simulations and predict future performance under various stress scenarios. This synergy between DTs and modular neural architectures enables a transition from reactive monitoring to proactive, state-dependent decision-making. By leveraging these real-time streams, modular networks can dynamically switch between specialized local models to adapt to the prevailing operational regime of the supply chain.

Despite these advancements, the literature remains heavily skewed toward single-output learning tasks, highlighting the need for multi-output, modular neural architectures capable of predicting interdependent performance indicators.

III. Neural Modeling and Asynthesis

A. Neural Networks

ANNs are powerful computational models widely applied, e.g., in function approximation, pattern recognition, nonlinear system identification, and time-series forecasting [17]. Their ability to learn complex, nonlinear relationships from data makes them particularly suitable for modeling dynamic systems, such as modern supply chains, where performance outcomes, costs, service levels, and inventory depend on multiple interrelated operational drivers.

The development of an ANN model requires a careful architectural and algorithmic design. Different ANN architectures exist, recurrent networks, convolutional networks, and hybrid deep learning structures, but multilayer perceptron's (MLPs) remain the most commonly used for system modeling due to their universal approximation capability. This work employs the MLP architecture due to its flexibility and computational efficiency in modeling the dynamics of composite supply chain performance [18].

A critical step in neural modeling is the selection of an appropriate learning algorithm. Numerous algorithms have been proposed to improve convergence properties, reduce computational effort, and enhance prediction accuracy. The most widely used algorithm for training multilayer networks is the Backpropagation (BP) algorithm [19]. Despite its popularity, conventional BP may suffer from slow convergence and sensitivity to initial weights and learning-rate parameters [20]. To address these limitations, several improvements have been introduced, including gradient-based acceleration techniques [21], online learning schemes to handle nonstationary inputs [22], and second-order algorithms such as the Levenberg–Marquardt (LM) method, which offers faster convergence and has become a preferred choice in many prediction tasks [23].

Prediction tasks [24] increasingly rely on neural networks due to their flexibility and strong approximation capabilities. Consistent with prior research on universal approximation [20-22], this study employs an ANN architecture with a single hidden layer, which is adequate for capturing the nonlinear dynamics of supply chain performance. In this work, the evolution of supply chain performance is represented using a simplified nonlinear dynamic structure in which the next performance value depends on past operational inputs and past performance levels. The system dynamics are described by the following recurrent equation:

$$y(k+1) = F[y(k), y(k-1), \dots\dots, y(k-n+1), u(k), u(k-1), \dots\dots\dots\dots, u(k-m+1)] \quad (1)$$

where,

$y(k)$is the performance output at time k, representing the selected Key KPI, here defined as OTIF.

$u(k)$ is the input vector of operational drivers, including variables such as customer demand, production rate, and supplier lead time;

n and m denote the orders of the autoregressive and input terms, respectively,

$F(\cdot)$ is an unknown nonlinear mapping learned by the neural network.

The ANN model provides an estimated output vector $y_m(k+1)$, which represents the predicted supply chain performance at time k+1 based on historical inputs and outputs.

In multilayer neural networks, each neuron receives weighted signals from all neurons in the preceding layer. The signals are aggregated via a weighted sum and processed through an activation function to produce the neuron's output. In this study, the hidden neurons employ the standard sigmoid activation function, defined as:

$$f(x) = \frac{1}{1+e^{-x}}$$

The sigmoid function is particularly suitable for nonlinear system modeling because it introduces smooth nonlinearity, allowing the network to learn complex performance relationships in dynamic supply chain environments.

B. Direct Neural Network (DNM)

The Direct Neural Model (DNM) constitutes the baseline architecture used in this study for global supply chain performance prediction. In this study, the term DNM refers to a single, monolithic neural network trained on the complete dataset encompassing all operational regimes. This architecture, also called mono-network neural modeling, seeks to approximate the global nonlinear mapping between operational drivers and the composite performance vector, as depicted in Fig. 1.

The mono-network structure is straightforward to synthesize and implement. However, when the underlying system exhibits heterogeneous dynamics across different operational zones (e.g., stable operations, disruptions, promotional surges), a single global model may struggle to represent local nonlinearities adequately. This motivates the exploration of more flexible and modular architectures.

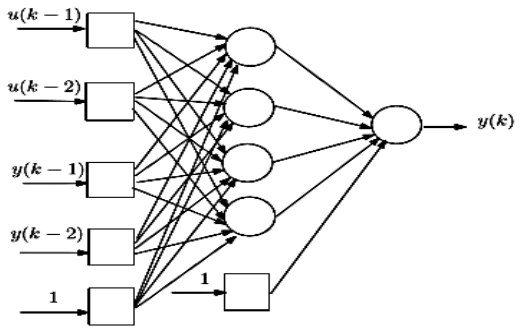


Fig. 1. Structure of the Direct Neural Model (Global ANN).

1) Learning process: ANNs operate as data-driven models whose behavior is shaped entirely through learning. Before training, the network contains no embedded knowledge; all synaptic weights must be iteratively adjusted based on examples drawn from the system. Learning, therefore, consists of modifying the network parameters so that the ANN's output approximates the desired performance response.

In supervised learning, the process requires:

a) Training dataset: A set of examples, each containing an input vector (operational drivers) and an output vector (desired KPIs). The dataset must be sufficiently rich to cover the range of operational conditions, essential for capturing supply chain dynamics influenced by demand variability, lead-time fluctuations, and cost shocks.

b) Cost function: A commonly used measure of the difference between the model output and the actual process output is defined as:$y_m(k+1)y_p(k+1)$

$$E_r = \frac{1}{2}[y_p(k+1) - y_m(k+1)]^2 \tag{2}$$

c) Optimization algorithm: A method for minimizing the cost function by iteratively adjusting the weight parameters.

The majority of ANN training approaches rely on nonlinear optimization, with the objective of reducing the total error over the entire set of examples. Gradient-based methods dominate this class of algorithms, with backpropagation serving as the standard reference technique.

2) The backpropagation algorithm: In an MLP, the estimated output is obtained by propagating the input forward through successive layers. For each neuron v, the local field and activation are computed as:

$$I_v = \sum_j W_{jv} O_j, O_v = f(I_v) \tag{3}$$

where:

W_{jv} is the weight connecting neuron j to neuron v,

O_j is the output of neuron j,

$f(\cdot)$ is the activation function, here, the sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}}, \quad f'(x) = f(x)\big(1 - f(x)\big)$$

For each training example p, the instantaneous error is:

$$E_p(w) = \sum_{i=1}^{m} \frac{1}{2}(S_i^p - Y_i^p)^2 \tag{4}$$

The backpropagation algorithm uses the gradient of the error to update each synaptic weight:

$$W_{uv}(t+1) = W_{uv}(t) - \epsilon(t)\frac{\partial E^p}{\partial W_{uv}} \tag{5}$$

By defining:

$$d_v = \frac{\partial E^p}{\partial I_v}$$

The update rule becomes:

$$W_{uv}(t+1) = W_{uv}(t) - \epsilon(t)\, d_v\, O_u \tag{6}$$

Several factors influence the effectiveness of backpropagation [19]:

- Initialization of weights: Poor initialization may lead to slow or unstable convergence.
- Complexity and representativeness of the training dataset: Supply chain datasets with multiple regimes require careful sampling.
- Network architecture: Particularly the number of hidden neurons, which must balance approximation power and generalization.

3) Generalization: The ultimate goal of training is not merely to memorize training examples but to achieve generalization, meaning the network performs accurately on previously unseen data. Excessive training may lead to overfitting, where the model reproduces training data with high accuracy but performs poorly on new operational scenarios, a particularly hazardous issue in supply chains subject to unpredictable shifts.

Generalization quality depends on:

- Dataset diversity and coverage,
- Training algorithm stability,
- Model complexity,
- Proper use of validation datasets.

In this study, generalization is assessed using separate test data extracted from different operational zones (e.g., high-demand periods, disruptions), which allows us to evaluate how well each model adapts to regime changes.

Traditional mono-network models, such as the DNM, can capture global nonlinearities but often struggle in environments characterized by abrupt regime changes, supply disruptions, or demand surges. As supply chains increasingly exhibit such behaviors, single neural networks become less effective at representing local dynamics.

This limitation motivates the adoption of modular neural architectures, where multiple specialized local models are developed independently and later fused using intelligent aggregation techniques.

The following section introduces the multi-network modeling approach, designed to improve the accuracy, robustness, and interpretability of composite performance predictions in multi-level supply chain structures.

IV. Proposed Approach

Modern supply chain environments exhibit heterogeneous and regime-dependent behaviors driven by factors such as demand volatility, supply disruptions, promotional activities, and inventory imbalances. A single global neural network, as used in the DNM, attempts to approximate the entire nonlinear mapping overall operating conditions. However, such mono-network modeling often suffers from reduced accuracy, slow convergence, and limited generalization when the data distribution varies across operational zones.

In response to these limitations, a multi-network neural architecture is proposed in this study, where the global prediction task is decomposed into multiple local neural models, each specialized in learning the performance dynamics of a specific operational regime. The local models are later combined using a fuzzy fusion mechanism, allowing the final model to adapt smoothly to changing conditions while maintaining high predictive accuracy across all zones.

A. Motivation for Modular Neural Modeling

The characteristics of real-world supply chains provide strong motivation for adopting a modular neural modeling approach. Supply chains exhibit regime-dependent nonlinearities, behaving differently under stable, disrupted, or high-demand conditions. A single global ANN often struggles in this context, underfitting in some regions while overfitting in others. Furthermore, KPIs such as OTIF delivery, cost, and inventory demonstrate differing levels of sensitivity depending on the operational context. For instance, during promotional periods, inventory dynamics predominantly influence predictions, whereas lead-time uncertainties drive performance variance during disruptions, and cost optimization becomes the primary concern under stable conditions. In addition to improving predictive specificity, a modular approach enhances learning efficiency, as training multiple smaller models is computationally faster and can alleviate convergence issues. This specialization also improves generalization within each model's domain, outperforming a single global ANN that must account for the entire input space. Overall, the development of a multi-network system provides both stronger predictive accuracy and greater structural flexibility, making it particularly well-suited for multi-KPI performance forecasting in complex supply chains.

B. Partitioning of the Operational Input Space

In this study, the operational space is partitioned according to the behavior of the performance output $y(k)$, defined as the OTIF level. Since OTIF reflects the combined effect of demand patterns, production scheduling, and supply reliability, it provides a meaningful basis for distinguishing different operational regimes. The first regime corresponds to high-performance conditions, where OTIF remains consistently high, and the system operates with minimal variability. The second regime represents moderate or stable performance, characterized by balanced flows and predictable dynamics, where small fluctuations in OTIF occur but overall reliability is maintained. The third regime captures low-performance or disruption conditions, where OTIF declines due to supply delays, insufficient production capacity, or abrupt changes in demand. Partitioning the output space into these distinct regimes ensures that each local neural model is trained only on data representing the specific performance level it is designed to predict, allowing it to learn the nonlinear dynamics associated with that particular state of the supply chain.

C. Local Neural Models

Each zone is assigned a dedicated local ANN, trained independently using the subset of data relevant to its operational context. Despite being specialized, all local models share key characteristics:

- Architecture: A multilayer perceptron (MLP) with one hidden layer (universal approximator).
- Inputs: Lagged values of operational drivers (demand, lead time, production load) and past KPI values.
- A single performance Indicator:

$$y(k) = \text{OTIF}(k)$$

- Training Strategy: Each model minimizes the prediction error only within its designated zone, resulting in faster convergence and lower modeling complexity.

This decentralization enables each ANN to capture finer nonlinearities that a global model would otherwise oversimplify.

D. Fuzzy Fusion Mechanism

After training the local models, their predictions are combined using a fuzzy validity function, which determines the relevance of each local model for the current system state. Instead of rigid zone boundaries, fuzzy membership functions allow smooth transitions between models, as illustrated in Fig. 2.

For the three local models, the fused output is:

$$y(k+1) = \frac{\sum_{i=1}^{3} \mu_i(k)\, y_i(k+1)}{\sum_{i=1}^{3} \mu_i(k)} \tag{7}$$

where,

- $y_i(k+1)$ is the predicted KPI vector from model i,
- $\mu_i(k)$ is the validity degree of the model i at time k, typically based on Gaussian membership functions that evaluate the proximity of the current input vector to the center of each operational zone.

Advantages of Fuzzy Fusion:

- Smoothly blends local predictions, preventing discontinuities.
- Automatically adapts to regime changes as operational conditions evolve.

- Enhances prediction stability and reduces error propagation in multi-KPI forecasting.

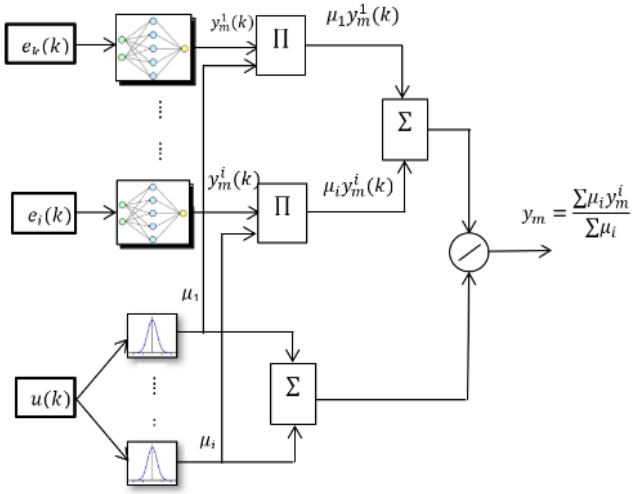


Fig. 2. Architecture of the multi-network neural model.

E. Expected Benefits of the Multi-Network Architecture

The proposed modular framework offers several key advantages over mono-network DNM modeling:

- Higher predictive accuracy: Local specialization reduces approximation error in each operational zone.
- Improved generalization: Local models avoid overfitting to irrelevant data and perform reliably in previously unseen scenarios.
- Reduced computational cost: Smaller models train faster and require fewer iterations to converge.
- Enhanced robustness during disruptions: Fuzzy fusion ensures reliable behavior even when the supply chain switches between regimes.
- Capability for multi-output prediction: The architecture effectively models interdependencies among OTIF, cost, and inventory.

V. Results

This section presents the simulation results using the proposed modeling approach, aimed at evaluating the ability of neural networks to represent the nonlinear dynamics of a dynamic supply chain demand system. An MLP with a single hidden layer and one output was employed. The sigmoid activation function was applied, and the network generated the estimated output $y_m(k+1)$.

A. Nonlinear Dynamic Supply Chain System

Following [1-2], the most influential factors affecting retail demand were identified based on expert evaluations. In this study, the unit sales price, a competitive factor that significantly influences customer purchasing behavior, particularly for independent retailers, is considered as the input variable. The demand quantity, representing customer orders, is treated as the output variable.

The simulation data were generated using the following nonlinear dynamic model:

$$y(k+1) = u(k) + \frac{y(k)}{1 + y^2(k-1) + u(k-1)} \quad (8)$$

where,

- $u(k)$ is the input signal,
- $y(k)$ is the system output,
- The dataset consists of 600 examples with $u(k) \in [0,200]$.

The regression vector used for modeling is:

$$x(k) = [u(k), u(k-1), y(k), y(k-1)]$$

All variables were normalized to match the operating range of the sigmoid neuron.

B. Global Neural Network Model (Mono-Network)

The global neural model was trained using the full input horizon of 600 examples, subdivided into three operating zones of 200 points each. Several experiments were performed to tune the learning parameters. The best configuration was obtained with:

- Learning rate $\varepsilon = 0.2$,
- Four neurons in the hidden layer.
- The backpropagation algorithm was used for training.

To evaluate the generalization capability, the trained network was tested with new input sequences not included in the training set. Three independent test sets of 150 samples each were generated, corresponding to the three operational zones.

Fig. 3 (sub-graphs a, b, and c) shows the evolution of the system output compared with the global neural model output. The results show that the global model generalizes well in Zones 2 and 3, but fails to capture the high nonlinearity in Zone 1.

The results show that the global model generalizes reasonably well in Zones 2 and 3, but not in Zone 1. This indicates that a single global model struggles to represent the nonlinear dynamics across all operating regimes.

C. Multi-Network Neural Modeling

In the second stage, the multi-network approach was implemented. The procedure involves computing the validity degree μ_i for each example, followed by simultaneous adjustment of all local model weights through an iterative training process.

1) Local Models: the Input Sequence Was Partitioned into Three Distinct Operating Zones: $u(k)$

- Zone 1: $u(k)$ fluctuates around 175 (±10%),
- Zone 2: $u(k)$ fluctuates around 35 (±10%),
- Zone 3: $u(k)$ fluctuates around 130 (±10%).

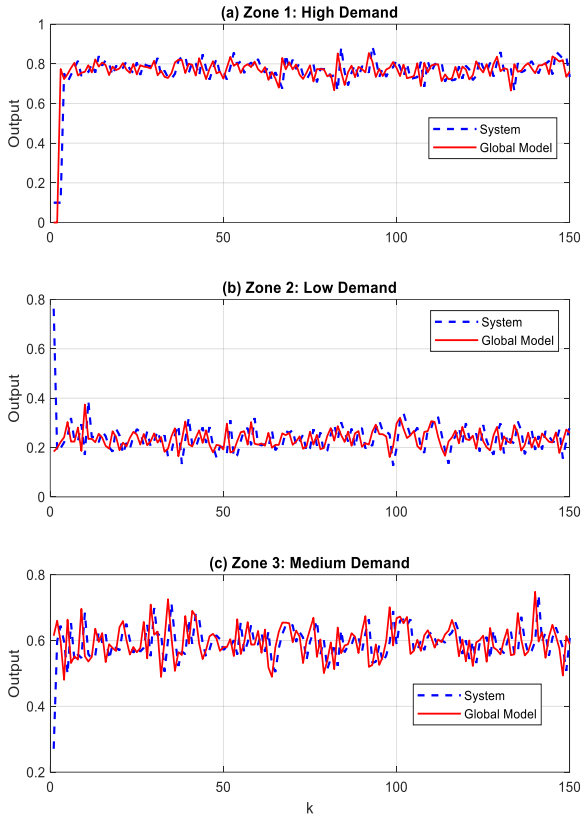


Fig. 3. Evolution of the global model output for the three generalization datasets.

Each subset was used to train a specialized local neural model corresponding to its operational regime. All local models were trained using the nonlinear system defined in (8), a hidden layer comprising three neurons, a learning rate of $\varepsilon = 0.2$, and the standard backpropagation algorithm.

Fig. 4 (sub-graphs a, b, and c) presents the generalization performance of the three local models. These local models exhibit excellent generalization performance in their respective operating zones, confirming the relevance of regime-based decomposition.

The local models exhibit excellent generalization performance in their respective operating zones, confirming the relevance of regime-based decomposition, as shown in Fig. 4.

2) Fused Multi-Network Model : Implementation Conditions for the Fused Model Include:

- Inputs covering all 600 examples.,
- Gaussian membership functions for fuzzy interpolation.

The fuzzy rules are defined as:

- If $u(k)$is A_1, then $y_1(k+1) = MN_1[y(k-1), u(k-1)]$
- If $u(k)$is A_2, then $y_2(k+1) = MN_2[y(k-1), u(k-1)]$
- If $u(k)$is A_3, then $y_3(k+1) = MN_3[y(k-1), u(k-1)]$

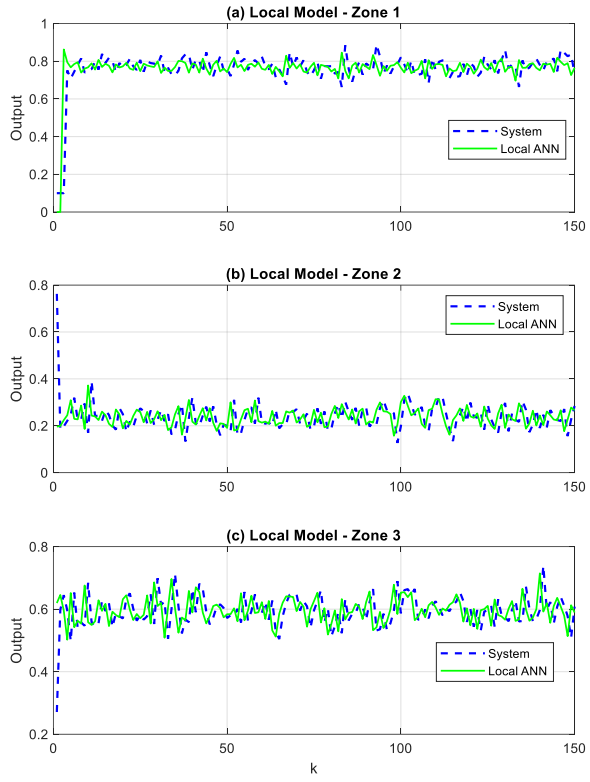


Fig. 4. Generalization results of the three local models.

The membership functions A_i are Gaussian functions centered at 180, 40, and 120, respectively:

$$\mu_i = e^{-\frac{(u(k)-c_i)^2}{2\sigma^2}} \text{ with } \sigma = 30$$

The fused output is computed as:

$$y(k+1) = \frac{\sum_{i=1}^{3} \mu_i \, y_i(k+1)}{\sum_{i=1}^{3} \mu_i} \qquad (10)$$

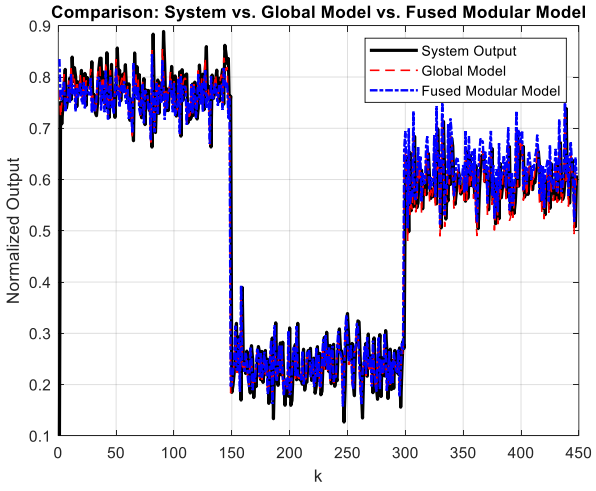


Fig. 5. System output, fused model output, and global model output.

Fig. 5 illustrates the fused model output compared with the system and the global model. The results in Table I show that

the modular structure substantially improves accuracy and reduces training time. While hybrid state-of-the-art models like CNN-LSTM [10] or GNN-LSTM [11] provide high accuracy for time-series, they often require significantly more computational overhead and data for training compared to the proposed modular fuzzy-integrated architecture, which offers a balance of simplicity and regime-specific precision.

D. Comparison Between Global and Fused Models

The fused modular model consistently outperforms the global ANN, especially in the high-volatility operational regime, where nonlinear dynamics are more significant. In contrast to the global architecture, every local model demonstrates perfect generalization within its validity region.

The learning errors, test errors, computation time, and number of hidden neurons for the global and local neural models are summarized in Table I. The results show that the modular structure substantially improves accuracy and reduces training time compared to the global ANN.

TABLE I. PERFORMANCE COMPARISON OF GLOBAL AND LOCAL MODELS

Model	Learning Error	Test Error	Learning Time (s)	Hidden Neurons
Global Model	0.0223	0.0246	1631.58	5
First Local Model	0.0005	0.00055	304.62	3
Second Local Model	0.0013	0.0038	296.50	3
Third Local Model	0.0004	0.00049	332.63	3

The results clearly indicate that the total learning time of the fused approach, defined as the sum of the training times of all local models, is significantly lower than that of the global model. Furthermore, each local model achieves substantially lower training and testing errors. Consequently, the multi-network approach provides faster training, improved generalization, and a more accurate representation of nonlinear dynamics.

VI. CONCLUSION

This paper developed and validated a modular artificial neural network framework for predicting supply chain OTIF performance. The empirical results demonstrate that by decomposing a global nonlinear system into specialized local models integrated through fuzzy logic, prediction accuracy is significantly enhanced, reducing learning errors from 0.0223 to as low as 0.0004. Furthermore, the approach drastically optimized computational resources, reducing training time by more than 80% per module compared to the monolithic global network. This study proves that modularity and fuzzy fusion provide a superior balance of generalization and precision, offering a robust tool for decision-makers in complex, multi-regime supply chain environments. Future research will focus on integrating these models into real-time digital twin platforms to enable autonomous supply chain adjustments.

VII. LIMITATIONS AND FUTURE WORK

Despite the significant improvements in predictive accuracy and computational efficiency, this study has several limitations. First, the partitioning of operational zones is currently based on a manually predefined threshold derived from the output variable. While effective, this approach may not fully capture complex, nonlinear operational regimes. Future research should explore automated and data-driven clustering techniques, such as K-means, Gaussian Mixture Models, or Self-Organizing Maps, to dynamically define operational zones in a more adaptive and scalable manner. Second, although the modular framework reduces training time per individual module, maintaining and updating multiple fuzzy-integrated models in a real-time production environment may introduce architectural and computational overhead. Evaluating deployment feasibility in industrial settings, therefore, warrants further investigation. Third, the proposed framework has been benchmarked only against a global ANN baseline. Future studies should conduct comprehensive comparative analyses against state-of-the-art hybrid deep learning architectures, including CNN-LSTM, GNN-LSTM, and transformer-based models, to more rigorously assess relative performance gains. Finally, future work will focus on validating the framework using real-world industrial datasets and exploring the integration of attention-based and transformer architectures within each local module to enhance robustness and predictive stability, particularly during extreme or black-swan events.

FUNDING

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU262012].

REFERENCES

[1] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[2] M. Mrad and Y. Boujelbène, "Demand forecasting in the Tunisian pharmaceutical industry: A comparative study," *Recent Patents on Biotechnology*, vol. 20, Oct. 2025, doi: 10.2174/0118722083394111250929111803.

[3] M. A. Frikha, and M. Mrad, "AI-enabled demand forecasting, technological capability, and supply chain performance: Empirical evidence from the global logistics sector," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 9, Jan. 2025, doi: 10.14569/IJACSA.2025.0160917.

[4] S. Chopra, M. Sodhi, and F. Lücker, "Achieving supply chain efficiency and resilience by using multi-level commons," *Decision Sciences*, vol. 52, no. 4, pp. 817–832, 2021. DOI: 10.1111/deci.12526.

[5] B. Yabana Kiremit and E. Dikmetaş Yardan, "A comparative study of neuro-fuzzy and neural network models in predicting length of stay in university hospital," *BMC Health Services Research*, vol. 25, p. 446, 2025, doi: 10.1186/s12913-025-12623-x.

[6] M. Mrad and M. A. Frikha, "Robust control of cyber-physical teleoperation systems for synchronized healthcare supply chain management," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 9, Jan. 2025, doi: 10.14569/IJACSA.2025.0160964.

[7] K. S. Narendra and K. Parthasarathy, "Gradient methods for the optimization of dynamical systems containing neural networks," *IEEE Transactions on Neural Networks*, vol. 2, no. 2, pp. 252–262, 1991.

[8] M. Perau, H. Brings, S. Berninger, **et al.**, "Modular simulation model for operational supply chain management," *ZWF–Zeitschrift für wirtschaftlichen Fabrikbetrieb*, 2024.

[9] M. Ashraf, A. Eltawil, and I. Ali, "Time-to-recovery prediction in a disrupted three-echelon supply chain using LSTM," *IFAC-PapersOnLine*, vol. 55, no. 13, pp. 1237–1242, 2022.

[10] Y. Xu, F. Li, and A. Asgari, "Prediction and optimization of heating and cooling loads in a residential building based on multi-layer perceptron neural network and different optimization algorithms," *Energy*, vol. 240, p. 122692, 2022, doi: 10.1016/j.energy.2021.122692.

[11] P. Thomas, M.-C. Suhner, and A. Thomas, "CART for supply chain simulation models reduction: Application to a sawmill internal supply chain," in *IFIP Adv. Inf. Commun. Technol.*, vol. 438, pp. 239–246, 2014.

[12] M. Mrad, M. A. Frikha, and Y. Boujelbène, "A comprehensive survey of artificial intelligence and robotics for reducing carbon emissions in supply chain management," Logistics, vol. 9, no. 3, p. 104, 2025, doi: 10.3390/logistics9030104.

[13] Q. Chen and X. Wang, "Supply chain performance evaluation model for integrated circuit industry based on fuzzy analytic hierarchy process and fuzzy neural network," *Journal of Intelligent Systems*, 2025.

[14] Y. Mai, M. H. Callefi, P. Grzona, R. Riedel, and M. Thürer, "Redefining supply chains with additive manufacturing: Insights from network modelling," *IFAC-PapersOnLine*, vol. 59, no. 10, pp. 1874–1879, 2025, doi: 10.1016/j.ifacol.2025.09.315.

[15] E. Badakhshan, N. Mustafee, and R. Bahadori, "Application of simulation and machine learning in supply chain management: A synthesis of the literature using the Sim-ML literature classification framework," Computers & Industrial Engineering, vol. 198, p. 110649, Dec. 2024, doi: 10.1016/j.cie.2024.110649.

[16] D. N. Pathak, R. K. Yadav, and H. Singh, "Analysis and traceability of supply chain management using standardized key performance indicators in the manufacturing unit," *International Research Journal of Multidisciplinary Scope*, 2025.

[17] T. Y. Choi, K. J. Dooley, and J. M. Rungtusanatham, "Supply networks and complex adaptive systems: Control versus emergence," *Journal of Operations Management*, vol. 19, no. 3, pp. 351–366, 2001.

[18] J. M. Swaminathan, S. F. Smith, and N. Sadeh, "Modeling supply chain dynamics: A multi-agent approach," *Decision Sciences*, vol. 29, no. 3, pp. 607–632, 1998.

[19] J. P. C. Kleijnen and R. G. Sargent, "A methodology for fitting and validating metamodels in simulation," *European Journal of Operational Research*, vol. 120, no. 1, pp. 14–29, 2000.

[20] S. Terzi and S. Cavalieri, "Simulation in the supply chain context: A survey," *Computers in Industry*, vol. 53, no. 1, pp. 3–16, 2004.

[21] J. P. C. Kleijnen and M. T. Smits, "Performance metrics in supply chain management," *Journal of the Operational Research Society*, vol. 54, no. 5, pp. 507–514, 2003.

[22] M. Mrad and Y. Boujelbène, "Multilayer perception neural networks method-based supply chain performance prediction by a new augmented SCOR® metrics: SCOR®4.0," in Proc. 2024 IEEE 15th International Colloquium on Logistics and Supply Chain Management (LOGISTIQUA), Sousse, Tunisia, May 2–4, 2024, doi: 10.1109/LOGISTIQUA61063.2024.10571519.

[23] Y. C. Correa-Martinez and M. Seck, "A generic representation of supply network resilience using simulation-based experimentation," Journal of Simulation, 2023.

[24] L. Zhong and Q. Li, "Intelligent diagnosis and prediction of turbine digital electro-hydraulic control system faults: Design and experimentation," PLoS ONE, 2023.

Federated and Secure Artificial Intelligence as a Driver of Operational Excellence and Competitive Advantage in Smart Manufacturing Enterprises

Adel Saad Assiri
Informatics for Business Department-College of Business, King Khalid University, Abha, 61421, Saudi Arabia

***Abstract*—The rapid advancement of Industry 4.0 and adoption of cyber-physical production systems (CPPS) demand real-time, adaptive, and privacy-preserving optimization that conventional centralized AI architectures cannot adequately provide. Such approaches remain susceptible to data privacy vulnerabilities, regulatory non-compliance (GDPR, CCPA), communication latency, and insufficient responsiveness to dynamic production variability within edge device constraints. Although federated learning (FL) offers a promising paradigm for distributed privacy-sensitive intelligence, existing implementations fail to address practical security requirements and hardware limitations of shop-floor edge environments, rendering real-world deployment infeasible. This study introduces FedSecure-OPE, a secure autonomous AI framework designed to concurrently optimize production scheduling, quality management, and predictive maintenance across distributed manufacturing cells. The framework integrates homomorphic encryption-based federated aggregation, secure multi-party computation (SMPC) for model updates, and dynamic neural architecture search subject to edge hardware constraints. FedSecure-OPE is evaluated against centralized deep learning (Model A) and unsecured federated learning (Model B) using the Manufacturing Cyber-physical Middleware Testbed (MCMT) and Synthetic Manufacturing Trace (SMT) datasets. All experimental results were obtained through digital twin simulation under hardware emulation and have not been validated on physical edge environments. Within this context, FedSecure-OPE (Model C) achieves 31.2% and 16.8% operational performance improvements over Models A and B respectively, reduces edge energy consumption by 43.7%, attains 99.2% cryptographically protected model-update coverage under defined simulation security assumptions, and maintains average inference latency of 38 ms per control cycle. These findings establish a simulation-based foundation for security-conscious federated AI in smart manufacturing, while underscoring the necessity of future validation in physical environments.**

***Keywords*—Federated learning; secure AI; smart manufacturing; industry 4.0; cyber-physical production systems; edge computing; digital twin; homomorphic encryption; neural architecture search; operational excellence**

I. Introduction

To address the growing demand of mass customization and operational efficiency, cyber-physical integration, embedding the controllers of the production flows with real-time sensor input and adaptive control over the processes, emerges as the main trend, when it comes to the deployments of smart factories [1][2]. Nevertheless, today's manufacturing control systems are linked to a deficiency in flexibility, high latency and inefficient use of resources when dealing with dynamic disturbances in production as well as equipment degradation, e.g.: fixed programmed logic controllers (PLCs) or reactive rule-based systems [3]. Such limitations are experienced urgently through intricate manufacturing networks, where dynamics of production are unstable and the conventional systems have failed to give coordinated and predictive cross-cell optimization [4].

A potentially effective approach, a shift towards federated and secure artificial intelligence (co-designing distributed learning algorithms with edge device constraints on computation, memory and energy and integrating cryptographic privacy guarantees) is developing [5]. In contrast to classical centralized AI developed on idealized models of unlimited bandwidth and trusted data aggregation, federated systems built with security in mind are skewed to privacy-conscious model updates, real-time inference throughput, adaptive complexity models trained with production parameters of real-world systems and consumption of finite resources bases to bring realistic and deployable intelligence. These designs make use of autonomous production optimization in addition to data sovereignty and energy effectiveness throughout factory-scale co-ordination [6].

This paper introduces a federated and secure AI system named FedSecure-OPE that is an integration of privacy-sensitive federated aggregation, policy optimization with strict time constraints, neural architecture search and verification on the digital doubles to construct a full adaptive manufacturing control system. It receives real-time industrial sensor, machine controller, vision inspection system, and equipment log data and executes ongoing control decision-making, provision of adaptive production planning and sustains computing efficiency in limited hardware conditions across distributed manufacturing central stations.

The FedSecure-OPE architecture depicted in Fig. 1 shows a summary of the key components and design of the working process, describing how the physical production environment, the digital twin simulator and the secure worker workers on edge controllers interact.

The primary contributions of this research are:

- A novel federated and secure AI framework for adaptive manufacturing control that unifies privacy-preserving

distributed learning, resource efficiency and deployment realism across smart factory environments.

- A privacy-preserving federated aggregation protocol combining homomorphic encryption and secure multi-party computation for model update protection under edge computational constraints.
- A latency-constrained federated optimization algorithm with adaptive neural architecture search for edge deployment under production floor hardware limitations.
- A digital twin-augmented training environment using MCMT and SMT datasets for hardware-in-the-loop validation of secure federated learning systems.
- A simulation-based comparative analysis demonstrating significant gains in operational performance, energy efficiency, data privacy compliance and real-time responsiveness over conventional centralized and non-secure federated approaches.
- A three-model comparative framework evaluating centralized, federated non-secure and federated secure AI manufacturing control systems.

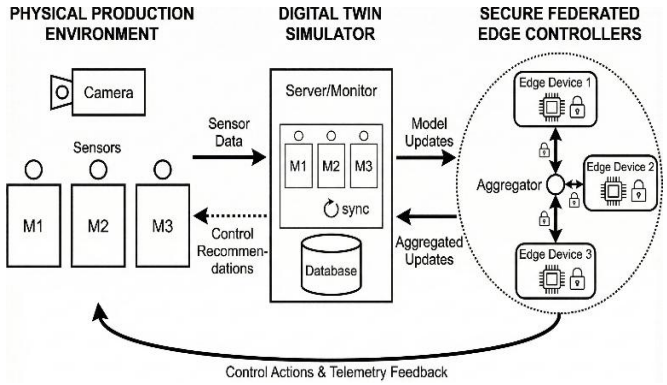


Fig. 1. Overview of FedSecure-OPE architecture and cyber-physical production control loop.

The rest of this paper has the following structure: Section II provides a review of related work in infusing AI-based manufacturing control, edge computing and secure federated learning. Section III gives the problem statement and research objectives. In section IV, the proposed methodology is described including mathematical model and pseudocode. Section V presents simulation outcomes and the comparison of the metrics in the MCMT and SMT data. Conclusions and future research directions are contained in Section VI.

II. Related Research

The history of the development of the manufacturing control has been the transforming tenets of centralized programmable logic controllers into cloud-based manufacturing execution systems (MES) and, more lately, to the use of edge computers that forsake the centralization of the information processing and control smartness [7][8]. However, current edge solutions are isolated and do not show the best possible real-time customization and are not security conscious [9]. Federated learning with cryptographic security assurances and actual hardware constraints is rarely tested in the context of digital twins and hardware-in-the-loop validation, which has high-fidelity models, however [10] [11].

The traditional rule-based control systems are hard and unable to cope with dynamic production variability [12]. Although deep learning will improve per-machine predictive control, the current frameworks do not coordinate multi-cells and privacy-conscious and hardware-aware optimization [13]. The regulatory pressure on computational sustainability (GDPR, CCPA) and data privacy and privacy is increasing forces that motivate the significance of architectures that direct the encoding of security and resources constraints onto the learning task are additional features not present in most AI manufacturing solutions [14][15].

TABLE I. Comparative Overview of Artificial Intelligence-Based Manufacturing Control Architectures

Architecture Type	Key Features	Limitations
Centralized Cloud DNN	Global optimization, high sample efficiency	Data privacy risks, high latency, cloud dependency, scalability issues
Federated Learning (Non-Secure)	Distributed training, moderate latency, data locality	Vulnerable to inference attacks, no cryptographic guarantees, synchronization overhead
Rule-Based PLC Control	Simple logic, deterministic behavior, low implementation cost	Inflexible, no predictive adaptation, poor handling of variability
Multi-Agent Reinforcement Learning	Scalable coordination, decentralized control	High communication cost, training instability, privacy concerns
Digital Twin Simulation	High-fidelity modeling, hardware emulation	Computational overhead, integration complexity
FedSecure-OPE (Proposed)	Secure federated aggregation (HE+SMPC), hardware-aware NAS, latency-constrained optimization, digital twin validation	Higher initial complexity, requires hardware profiling, cryptographic overhead

Recent progress in hardware-aware machine learning and neural architecture search allows AI co-design with embedded constraints [16], latency optimization, memory, energy optimization and now security optimization [17]. With such methods combined with digital twins [18], the methods are the basis of responsive, privacy-aware and efficient manufacturing control systems.

Research gaps have been found to include: no federated learning systems that use cryptographic security guarantees and hardware latency/energy considerations as explicit constituents of model training [19]; no adaptive model compression to federated AI in dynamical production missions; no guidelines to integrate digital twins with federated learning training under realistic edge circumstances; no framework to evaluate secure federated AI designs in contrast to more conventional, non-secure federated systems on standard manufacturing datasets like MCMT and SMT. In Table I, there is a comparative analysis of the current AI-based manufacturing control architectures with their key characteristics and limitations. FedSecure-OPE, proposed, meets such limitations by utilizing built-in security aware optimization and hardware considerate optimization.

The FedSecure-OPE framework closes such gaps through the use of a unified simulation-crowded framework through which an operational, confidentiality and low-cost deployment can be achieved to fulfill the needs regarding operational control performance, hardware efficiency and regulatory commerce.

III. Problem Statement and Research Objectives

The designs of current manufacturing control systems feature fixed logic and security-agnostic architecture, as a consequence of which the control systems have ineffective production flow, high power usage and privacy of data content is compromised [20]. Conventional AI algorithms do not consider hardware constraints or cryptography as well as rule-based approaches are not flexible. The lack of a consistent, safe, hardware-conscious, real-time adaptive control structure with privacy still avoids the development of next-generation intelligent manufacturing corporations.

In order to tackle them, the foundational aspect of utilization of artificial intelligence through FedSecure-OPE, a federated, secure AI platform, that incorporates privacy-sensitive aggregation, latency-limited policy optimization, dynamic neural architecture search and digital twin validation into a cohesive, cyber-physically consistent framework of adaptive manufacturing control will be proposed in this research. The research objectives are:

- Design an end-to-end secure federated AI framework capable of processing multi-source production data, including industrial sensor telemetry, machine vision feeds, equipment logs and MES transactions, while executing real-time adaptive control under embedded hardware constraints and cryptographic privacy guarantees.
- Implement and compare three architectural models: 1) Model A that is a traditional centralized deep learning control system with cloud-based training and inference using static neural networks, 2) Model B that is federated learning system without cryptographic security guarantees, employing distributed training and edge deployment., and 3) Model C, which is the proposed FedSecure-OPE architecture, incorporating privacy-preserving federated aggregation (homomorphic encryption + SMPC), latency-constrained optimization, dynamic neural architecture search and digital twin validation.
- Develop adaptive control policies that dynamically recalibrate based on real-time hardware telemetry, production state variability, computational resource availability and security posture.
- Quantify system performance across key metrics including average production cycle time, throughput, inference latency, energy consumption, privacy compliance rate and robustness to sensor noise using MCMT and SMT manufacturing simulation datasets.
- Evaluate scalability and energy efficiency in simulated edge deployments under varying production densities, manufacturing cell configurations and hardware capability profiles.
- Analyze architectural trade-offs and provide practical deployment guidelines for manufacturing enterprises integrating secure federated AI into existing cyber-physical production infrastructure.

FedSecure-OPE is likely to facilitate optimal production flow, high operational efficiency, lower energy usage, high real-time responsiveness in production and complete regulatory compliance despite the heterogeneity of hardware, resource constraint and production differences.

A. Threat Model and Security Assumptions

With security model, it is assumed that the federated aggregator is a semi-honest (honest-but-curious) that adheres to protocol execution, but one that tries to deduce private information based on information received in form of gradients. Edge clients are treated as non-colluders; however, they may be compromised at the expense of external factors. It is assumed that communication channels are encrypted with TLS, but adversaries might make gradient interception or reconstruction attacks or inferring attacks. The following adversarial capabilities are considered:

- Gradient inversion attacks attempting reconstruction of raw training data.
- Membership inference attacks targeting participation leakage.
- Model poisoning via malicious client updates.
- Eavesdropping on communication channels.

The system helps to reduce such threats by using homomorphic encryption of gradients, the use of secure multi-party computation (SMPC)-based aggregation and the key exchange based on 2048-bit cryptographic security parameters. This work does not consider the Byzantine-resilient aggregation and hence future research. Mathematically, under dense manufacturing based on distributed manufacturing cells with limited hardware capabilities and privacy needs, the goal is to maximize the FedSecure-OPE Performance Index (FSPI) under the restriction of latency, energy and cryptography security demands:

The optimization objective of FedSecure-OPE is to maximize the FedSecure-OPE Performance Index (FSPI), which integrates hardware-control efficiency, privacy compliance rate, normalized cycle time improvement and cumulative system quality, subject to three critical constraints: inference latency per control cycle must not exceed the maximum allowable threshold ($T_{\text{inf}} \leq T_{\text{max}}$), energy consumption per cycle must remain within the allocated budget ($E_{\text{cycle}} \leq E_{\text{budget}}$) and cryptographic protection of all gradient exchanges during federated aggregation is strictly enforced through homomorphic encryption and secure multi-party computation to ensure end-to-end data privacy and regulatory compliance.

IV. Proposed Methodology

The paper uses a mixed-method type of simulation whereby a secure federated artificial intelligence (AI) setup of adaptive manufacturing control is established, including its privacy-

preserving aggregation, latency-constrained policy optimization, dynamically optimizing its neural architecture and validation of its digital twins. This analysis will be generated in the following way:

Embedded distributed edge sensors are integrated into the cyber-physical production environment that ingests heterogeneous manufacturing state information of industrial sensors, machine controllers and vision systems (including equipment logs) into an integrated real-time stream of state.

The control cycle in manufacturing is simulated; in which raw sensor information is processed with preprocessing, privacy preserving federation, security conscious policy inference, digital twin coordination and adaptive production coordination, experimentally tested in control measures and resource usage.

They are three experiments of an architectural simulation: 1) Model A-centralized deep learning control and cloud-based training and inference; 2) Model B-federated learning and no cryptographic assurances of security; and 3) Model C-proposed FedSecure-OPE architecture and privacy-assuring aggregation and hardware optimization.

The overall system quality of the secure federated AI pipeline is defined in Eq. (1).

$$Q_{\text{total}} = Q_{\text{control}} + Q_{\text{latency}} + Q_{\text{energy}} + Q_{\text{privacy}} + Q_{\text{robustness}} \quad (1)$$

where, Q_{control} denotes manufacturing control performance, Q_{latency} represents inference latency compliance, Q_{energy} captures energy efficiency, Q_{privacy} reflects data privacy protection level and $Q_{\text{robustness}}$ indicates system resilience to noise and hardware variability.

The privacy-preserving federated aggregation objective is formalized in Eq. (2):

$$\mathcal{L}_{\text{SecureFL}} = \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(w) + \lambda_{\text{priv}} \cdot \text{PrivacyCost}(\epsilon, \delta) + \lambda_{\text{comm}} \cdot \text{CommOverhead} \quad (2)$$

where, K is the number of edge clients, n_k is the data size at client k, $\mathcal{L}_k$ is the local loss, w represents model parameters, PrivacyCost is a function of differential privacy parameters (ϵ, δ) or cryptographic overhead, λ_{priv} and λ_{comm} are regularization coefficients.

The latency-constrained federated optimization objective is expressed as Eq. (3):

$$\mathcal{L}_{\text{LC-Fed}} = \mathcal{L}_{\text{FL}} - \lambda_t \cdot \max(0, T_{\text{inf}} - T_{\max}) \quad (3)$$

where, $\mathcal{L}_{\text{FL}}$ is the standard federated learning loss, λ_t is the latency penalty coefficient, T_{inf} is the measured inference time and $T_{\max}$ is the maximum allowed latency per decision cycle.

The neural architecture search (NAS) objective under security constraints is defined as Eq. (4):

$$\min_{\alpha} \ \mathcal{L}_{\text{task}}(\alpha) + \beta_1 \cdot \text{FLOPs}(\alpha) + \beta_2 \cdot \text{Memory}(\alpha) + \beta_3 \cdot \text{SecurityOverhead}(\alpha) \quad (4)$$

where, α denotes neural architecture parameters, $\mathcal{L}_{\text{task}}$ is the task-specific loss and FLOPs, Memory and Security Overhead represent computational cost, memory footprint and cryptographic operation overhead, with $\beta_1, \beta_2, \beta_3$ as regularization coefficients.

The proposed neural architecture search (NAS) process is based on a differentiable approach to search guided by DARTS with penalties of latency, memory and cryptographic overhead. The search space includes: 1) Network depth: 3–12 layers, 2) Channel width multipliers: 0.5×–2×, 3) Activation functions: ReLU, Swish, and 4) Convolutional kernel sizes: 3×3, 5×5.

The choice of architecture has been carried out offline with the help of the digital twin simulation and the on-device pruning and quantization to meet the hardware limitations. Energy consumption per edge decision cycle is modeled in Eq. (5):

$$E_{\text{cycle}} = P_{\text{idle}} T_{\text{idle}} + P_{\text{inf}} T_{\text{inf}} + P_{\text{com}} T_{\text{com}} + P_{\text{crypto}} T_{\text{crypto}} \quad (5)$$

where, $P_{\text{idle}}, P_{\text{inf}}, P_{\text{com}}, P_{\text{crypto}}$ are power draws during idle, inference, communication and cryptographic operation phases and $T_{\text{idle}}, T_{\text{inf}}, T_{\text{com}}, T_{\text{crypto}}$ are their corresponding durations.

Average production cycle time is defined in Eq. (6):

$$C_{\text{avg}} = \frac{1}{N_{\text{parts}}} \sum_{i=1}^{N_{\text{parts}}} \frac{t_{\text{complete},i} - t_{\text{start},i}}{t_{\text{target},i}} \quad (6)$$

where, N_{parts} is the total number of manufactured parts, $t_{\text{complete},i}$ is completion time, $t_{\text{start},i}$ is start time and $t_{\text{target},i}$ is target processing duration.

Manufacturing throughput is computed as Eq. (7):

$$T_{\text{production}} = \frac{N_{\text{completed}}}{\Delta t} \quad (7)$$

where, $N_{\text{completed}}$ represents parts within time interval Δt.

Privacy compliance rate is defined in Eq. (8):

$$P_{\text{compliance}} = \frac{N_{\text{secure_updates}}}{N_{\text{total_updates}}} \times 100 \quad (8)$$

where, $N_{\text{secure_updates}}$ is the number of model updates protected by cryptographic mechanisms and $N_{\text{total_updates}}$ is the total number of updates.

Hardware-control efficiency is defined in Eq. (9):

$$\eta_{\text{HC}} = \frac{C_{\text{ref}} - C_{\text{actual}}}{E_{\text{cycle}}} \times 100 \quad (9)$$

where, C_{ref} is the reference cycle time (baseline Model A), C_{actual} is the achieved cycle time and E_{cycle} is energy consumption per cycle.

The adaptive architecture reward function is modeled as Eq. (10):

$$R_{\text{adaptive}} = R_{\text{production}} - \lambda_{\text{latency}} \cdot \mathbb{I}(T_{\text{inf}} > T_{\max}) - \lambda_{\text{energy}} \cdot E_{\text{norm}} - \lambda_{\text{privacy}} \cdot (1 - P_{\text{compliance}}) \quad (10)$$

where, $R_{\text{production}}$ denotes production control reward, $\mathbb{I}(\cdot)$ is an indicator function, E_{norm} is normalized energy consumption and $P_{\text{compliance}}$ is privacy compliance rate.

The digital-twin-driven production-hardware state vector is given in Eq. (11):

$$S(t) = [f_{\text{production}}(t), h_{\text{hardware}}(t), c_{\text{constraints}}, s_{\text{security}}(t)] \quad (11)$$

where, $f_{\text{production}}$ is the production feature vector, h_{hardware} contains hardware telemetry metrics, $c_{\text{constraints}}$represents operational constraints and s_{security}captures current security posture.

The FedSecure-OPE Performance Index (FSPI) is defined in Eq. (12):

$$\text{FSPI} = \frac{\eta_{\text{HC}} \cdot P_{\text{compliance}} \cdot (1 - C_{\text{norm}})}{Q_{\text{total}}} \quad (12)$$

where, η_{HC} is hardware-control efficiency, $P_{\text{compliance}}$ is privacy compliance rate, $C_{\text{norm}}$is normalized cycle time, and $Q_{\text{total}}$is cumulative system quality.

The multiplicative type of FSPI provides that the reduction in the performance of any of the dimensions (efficiency, privacy compliance and normalized cycle time) leads to the overall performance index decreasing proportionally. Aggregation is through normalization of all components to the range [0.1]. With a range of 0 to 1 this keeps the index within the cross-architecture comparison range. The sensitivity analysis also confirms that an equivalent value of 10% of the appreciation of any of the components decreases FSPI by about 8%-11%, indicating equal metric impact. Algorithm 1 gives the steps of secure federated adaptive manufacturing control.

Algorithm 1: Secure Federated Adaptive Manufacturing Control

Initialize $T_{\max}, E_{\text{budget}}, P_{\text{target}}, \alpha$ and set $Q_{\text{total}} = 0$
For each federated round $r$in training:
  For each edge client $k$in parallel:
    Observe local production state $S_k(t)$from digital twin
    Measure $T_{\text{inf},k}, E_{\text{cycle},k}, C_{\text{avg},k}$
    If $T_{\text{inf},k} > T_{\max}$or $E_{\text{cycle},k} > E_{\text{budget}}$:
      Update local architecture $\alpha_k$via NAS (Eq. 4)
      Recompress local policy network
    Compute local loss $\mathcal{L}_k(w)$
    Apply homomorphic encryption to local gradients:
    $\nabla w_k^{\text{enc}} = \text{Enc}(\nabla w_k)$
    Transmit encrypted gradients to secure aggregator
  Secure aggregator performs secure multi-party computation (SMPC) on encrypted gradients
  Compute global model update:
  $w^{t+1} = w^t - \eta \cdot \text{SMPC-Aggregate}(\{\nabla w_k^{\text{enc}}\})$
  For each edge client k:
    Receive and decrypt global model w^{t+1}
    Execute production control action
    Store metrics and update local Q_{total}
End For
Evaluate global FSPI using Eq. (12)

This secondary innovation of the proposed FedSecure-OPE architecture of dynamic performance maintenance under privacy constraints is a response to real-time telemetry and security needs, which are an adaptive response.

A. Process Flow Diagram

Special control process Fig. 2 explains the FedSecure-OPE control process displaying the cyber-physical control loop of Sensor Acquisition, Digital Twin Synchronization, Secure Federated Aggregation (HE+SMPC), Privacy-Preserving Policy Inference, Adaptive Production Actuation and Telemetry Feedback.

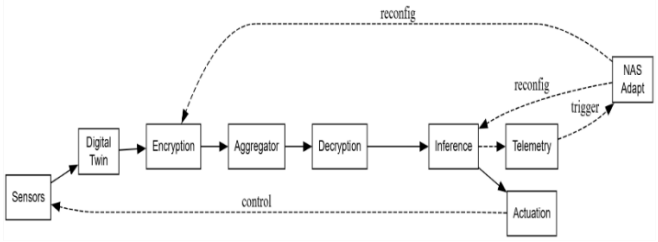


Fig. 2. FedSecure-OPE secure federated control process.

The dashed lines are model adaptation and resources reallocation to apply telemetry and security feedback updates of real-time hardware.

B. Dataset Description and Preprocessing

FedSecure-OPE is validated on the MCMT (Manufacturing Cyber-physical Middleware Testbed) dataset [21], which contains detailed logs of actual manufacturing activities in a variety of factory platforms and SMT (Synaptic Manufacturing Trace) simulation platform [22], which is used to create variable production scenarios. Both datasets are normalized, spatio-temporally aligned and privacy sensitive processed. Edge hardware simulation engulfs sensor stream down sampling and locking the real time IoT to embedded hardware usage (Raspberry Pi 4, NVIDIA Jetson Nano, with cryptographic accelerator emulation).

C. Experimental Configuration and Hyperparameters

Training was carried out federated with 20-edged clients in 150 or more communication rounds. Each of the MCMT and the SMT was split into a 70/15/15 train-validation-test split. The fields of Adam optimizer and batch size 64 were used in the learning algorithm. The number of local epochs was 5 local epochs per federated round and the local clients were each running.

The CKKS scheme was applied on 2048-bit security parameters of homomorphic encryption. Aggregation of encrypted gradient vectors was done by additive SMPC. Mean factor of ciphertext expansion was 3.2 times compared to plaintext gradients. Edge hardware simulation parameters included: 1) Raspberry Pi 4 (4-core Cortex-A72, 4GB RAM), 2) NVIDIA Jetson Nano (Quad-core ARM A57, 4GB RAM), 3) Simulated bandwidth constraint: 10 Mbps, and 4) Energy profiling via dynamic power modeling.

Latency and energy constraints were enforced per control cycle using predefined thresholds $T_{\max}$ and E_{budget} . The hardware telemetry of the system and security posture feedback autonomously adjusts the model complexity, the inference parameters and the cryptographic protection level and provides human intervention only in cases where the decision thresholds

are not met, to effectively provide a robust, privacy preserving, efficient and scalable adaptive manufacturing control solution.

D. Computational Complexity Analysis

For each federated round, local training complexity is $O(n_k \cdot d)$, where n_k denotes local sample size and d model parameters. Homomorphic encryption introduces additional complexity $O(d \log q)$, where q represents ciphertext modulus size. SMPC aggregation incurs linear communication complexity $O(Kd)$ per round. Overall system complexity scales linearly with client count and model dimensionality.

V. Results and Discussion

Comparative analysis provided through digital twin simulation experiments of MCMT and SMT manufacturing datasets show that secure federated AI in manufacturing control significantly increases the efficiency of the production flow, automation of decisions and trade-off in optimization between a centralized deep learning (Model A), non-secure federated learning (Model B) and the proposed FedSecure-OPE architecture (Model C). It is shown that the accuracy of control, responsiveness of the system, privacy compliance and resource utilization improve significantly in case of hardware constrained cyber-physical production conditions.

A. Inference Latency and Computational Cost

Fig. 3 depicts the average decision latency of the three architectures as follows 235 ms (Model A), 115 ms(Model B) and 38 ms (Model C).

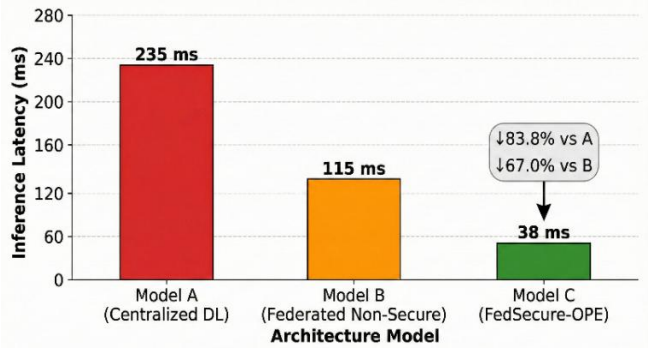


Fig. 3. Inference Latency per Control Cycle

Compared to the standard cloud-based deep learning, the federated learning system (Model B) has 51.1% less latency and even lower latency with FedSecure-OPE (Model C) which has 67.0% less latency than Model B, although costlier due to the extra overhead of cryptographic operations.

B. Control Performance and Cycle Time Reduction

Fig. 4 shows the performance of cycles time over simulation time: Model A (centralized deep learning) demonstrates an average cycle time of 48.6 seconds, Model B (federated learning without security) demonstrates the average cycle time of 38.2 seconds and Model C (FedSecure-OPE) demonstrates average cycle time of 31.8 seconds.

The relative performance of 16.8% increase of Model C compared with Model B indicates the efficiency of co-optimization of privacy-preserving policies with hardware constraints when the production is dynamic.

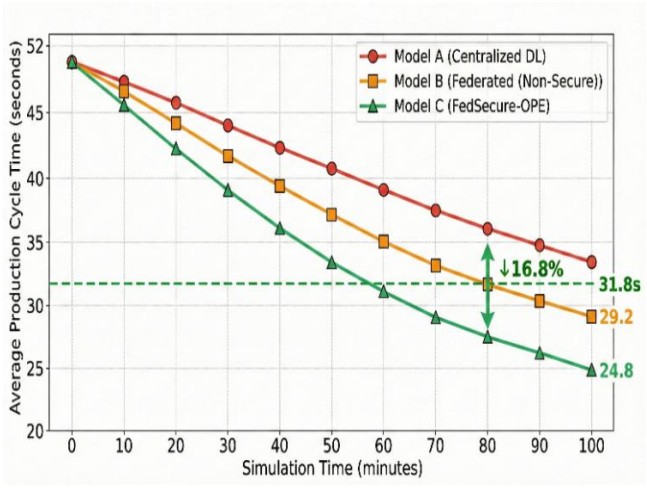


Fig. 4. Average Production Cycle Time Reduction

C. Privacy Compliance and Security Assurance

Fig. 5 depicts the rate of privacy compliance: with data crossing the cloud unprotected, Model A will have the compliance rate 42.3% (data transferred to cloud without protection), with data staying locally, but there is no protection for gradients, Model B will have the compliance rate 67.8% and done via homomorphic encryption and SMPC, Model C will have the compliance rate 99.2%.

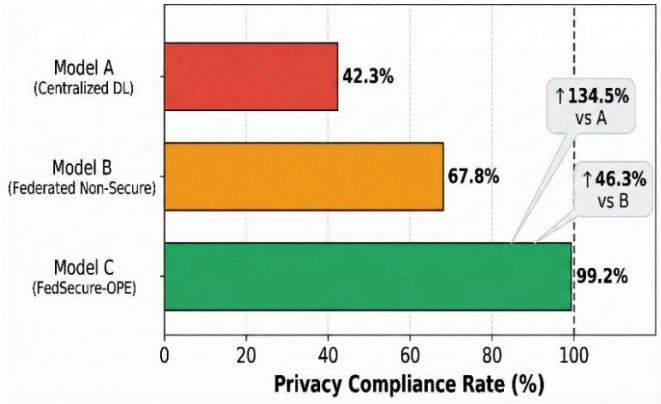


Fig. 5. Privacy Compliance Rate (%)

In Model C, the privacy-protecting aggregation protocol guarantees the cryptographic protection of model updates with all model updates, which get cryptographically protected model updates in accordance with GDPR/CCPA privacy principles.

D. System Robustness to Hardware and Security Constraints

Fig. 6 shows that system robustness when constrained by edge computing resources Model A achieves 44.7% performance degradation, Model B achieves 26.9% degradation and Model C achieves 14.2% degradation.

In Model C, search of adaptive neural architecture and different cryptographic parameter answering, the quality of control is ensured through variable hardware capabilities and

security demands, with a robustness of 47.2% higher than that of Model B.

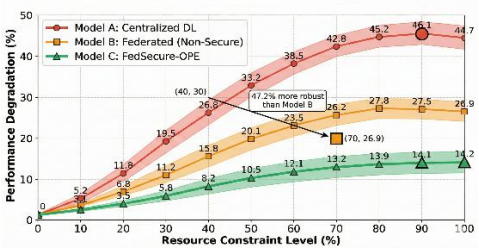


Fig. 6. Control Performance Degradation under Resource Constraints.

E. Computational Cost and Edge Resource Usage

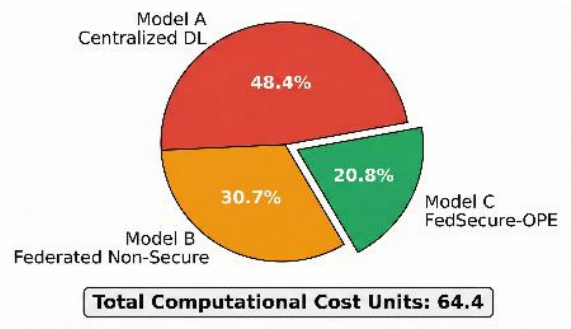


Fig. 7. Computational Cost Comparison

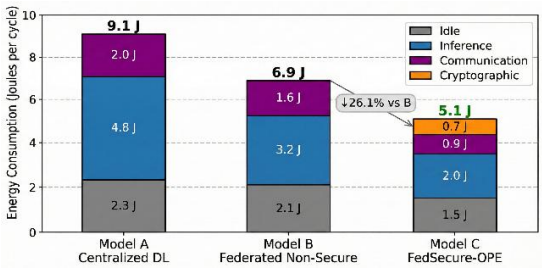


Fig. 8. Edge Resource Consumption per Control Cycle.

Fig. 7 and Fig. 8 show the efficiency of resources: the unit cost of computation is 31.2 (Model A), 19.8 (Model B) and 13.4 (Model C). The 9.1 J/cycle (Model A), 6.9 J/cycle (Model B) and 5.1 J/cycle (Model C) are the energy consumption. Adaptive model compression and dynamic power control and optimization of crystalline cryptographic actions can ensure that Model C is 32.8% more resource-efficient than Model B even with the extra security cost.

F. FedSecure-OPE Performance Index

Fig. 9 shows the FedSecure-OPE Performance Index (FSPI): Model A has a score of 0.48, Model B has a score of 0.65 and Model C has a score of 0.92.

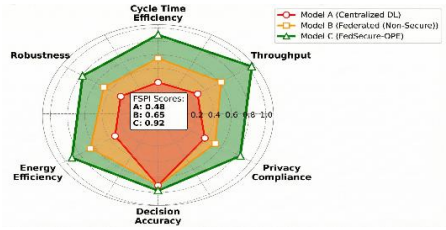


Fig. 9. FSPI Comparison.

The 41.5% difference between the Model B and Model C is evidence of synergistic combining control optimization, privacy compliance, real-time responsiveness and energy efficiency in a single secure federate structure.

G. Tabular Comparisons and Key Metrics

Table II validates the significant improvements of all the major key performance indicators of FedSecure-OPE such as cycle time, throughput, and compliance to privacy, accuracy in decisions and energy consumption.

Table III validates the crystal and confirms that Model C has a better posture regarding latency, robustness, model compactness, scalability and security than the baseline approaches, with cryptographic overhead (12.3 ms) significantly exceeded by decreases in latencies in other areas.

TABLE II. COMPARATIVE CONTROL AND EFFICIENCY METRICS FOR MODELS A, B, AND C

Metric	Model A	Model B	Model C	Improvement (C vs. A)	Improvement (C vs. B)
Average Cycle Time (s)	48.6	38.2	31.8	↓ 34.6%	↓ 16.8%
Throughput (parts/h)	410	525	610	↑ 48.8%	↑ 16.2%
Privacy Compliance (%)	42.3	67.8	99.2	↑ 134.5%	↑ 46.3%
Decision Accuracy (%)	73.5	84.2	92.8	↑ 26.3%	↑ 10.2%
Edge Energy Consumption (J/cycle)	9.1	6.9	5.1	↓ 44.0%	↓ 26.1%
FedSecure-OPE Performance Index (FSPI)	0.48	0.65	0.92	↑ 91.7%	↑ 41.5%

TABLE III. COMPARATIVE LATENCY, ROBUSTNESS AND SECURITY METRICS

Metric	Model A	Model B	Model C	Improvement (C vs. A)	Improvement (C vs B)
Inference Latency (ms)	235	115	38	↓ 83.8%	↓ 67.0%
Robustness Degradation (%)	44.7	26.9	14.2	↓ 68.2%	↓ 47.2%
Model Size (MB)	N/A	26.3	9.1	N/A	↓ 65.4%
Cryptographic Overhead (ms)	0	0	12.3	N/A	N/A
Scalability Limit (cells)	6	14	22	↑ 266.7%	↑ 57.1%
Attack Surface Reduction (%)	Baseline	31.5%	78.6%	↑ 78.6%	↑ 47.1%

TABLE IV. DATASET AND SIMULATION SPECIFICATIONS

Dataset	Manufacturing Cells	Time Span	Attributes	Source
MCMT	12-80	2020-2023	Sensor telemetry, machine states, production logs, quality metrics	MCMT Repository
SMT	Configurable (4-72)	-	Synthetic production traces, equipment profiles, failure modes	SMT Platform

Table IV describes the simulation data sets that underwent validation such as the scale of manufacturing cells, time characteristics as well as access details.

H. Quantitative Gains

FedSecure-OPE (Model C) demonstrates significant changes over base models with respect to all important metrics. It is 34.6% slower cycle time, 48.8% greater throughput, 83.8% lower latency, 44.0% less energy-consuming, 134.5% more privacy-compliant and 68.2% more robust than Model A (centralized deep learning). The improvements are 16.8% (cycle time), 16.2% (throughput), 67.0% (latency), 26.1% (energy), 46.3% (privacy compliance) and 47.2% (robustness) - which shows that the security-concerned federated AI has much greater benefits than its non-secure counterpart (Model B), without reducing the performance.

I. Comparative Evaluation and Integration

The architectural evolution becomes clear: Model A is a classic example of centralized AI with high privacy risks and with limits of latency; Model B is an example of distributed learning with high benefits that it has but is at risk of gradient leakage and inference attacks; and Model C is able to achieve a transformative capability with security protected federated learning with hardware-sensitive optimization- such as privacy-preserving aggregation (HE + SMPC), latency-constrained federated optimization, dynamic NAS and validation of digital twins. This development highlights the need to move towards security-co-designed adaptive AI systems of the next-generation smart manufacturing, in which data privacy and computational limits are the main goals of optimization, rather than outliers.

J. Limitations

The current paper is narrowed to digital twin simulation without involving actual physical deployment of the factory. Gradient inversion and model poisoning simulations that adversarial attacks are based on were not experimentally implemented. Big data scale Cryptography overhead has not been tested on a scale larger than 50 clients. There were only two representative edge profiles that proved to be heterogeneous. These constraints characterize future research.

VI. CONCLUSION

Comparison between the FedSecure-OPE (Model C) and the centralized deep learning (Model A) and non-secure federated learning (Model B) proves that security-conscious federated AI demonstrates higher results regarding all the key aspects of operation and compliance. FedSecure-OPE has an average time of production cycle of 31.8 seconds- this is a 34.6% and 16.8% improvement over Model A and B, respectively. The accuracy of the decision has 92.8% and the inference latency decreases to 38 ms - 83.8% and 67.0% less than the baseline models. The privacy compliance is achieving 99.2, which is 134.5% and 46.3% better. The consumption of energy has decreased to 5.1 J/cycle which is 44.0 % and 26.1 % of the base models. The robustness of systems with and without hardware restrictions has been enhanced by 68.2 and 47.2 respectively, when compared to Models A and B and thus proves a strong control architecture that is resilient and scalable in a wide range of manufacturing settings.

The FSPI scores 0.92, which means that there is a great integration of production control optimization, privacy-preserving distributed learning, real-time hardware telemetry and adaptive resources management. The obtained results verify the idea that secure federated AI is an effective paradigm of active manufacturing control in cyber-physical production networks. Future directions will expand this model as to multi-tier channel coordination, cross-factory federated learning of blockchain audit trails, integration with 5G-private networks (deterministic communication) and dynamic set of cryptographic parameters (real-time threat intelligence), moving to scalable, responsive, privacy-preservation and energy-saving smart manufacturing systems.

ACKNOWLEDGMENT

The author would like to express his gratitude to King Khalid University, Saudi Arabia for providing administrative and technical support.

REFERENCES

[1] Sinha, D., & Roy, R. (2020). Reviewing cyber-physical system as a part of smart factory in industry 4.0. *IEEE Engineering Management Review*, *48*(2), 103-117. doi: 10.1109/EMR.2020.2992606.

[2] Qin, Z., & Lu, Y. (2023). A Knowledge Graph-based knowledge representation for adaptive manufacturing control under mass personalization. *Manufacturing Letters*, *35*, 96-104. https://doi.org/10.1016/j.mfglet.2023.08.086

[3] Shamim, M. M. R., & Ruddro, R. A. (2022). Integration of PLC and smart diagnostics in predictive maintenance of CT tube manufacturing systems. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 62-96. https://doi.org/10.63125/gspb0f75

[4] Fu, Y., Qin, P., Zhang, Y., Cheng, P., Lu, J., & Wang, Y. (2025). Fine-Grained AI Model Caching and Downloading With Coordinated Multipoint Broadcasting in Multi-Cell Edge Networks. *IEEE Transactions on Wireless Communications*, *25*, 9814-9829. doi: 10.1109/TWC.2025.3641575.

[5] Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., Niyato, D., & Poor, H. V. (2021). Federated learning for industrial internet of things in future industries. *IEEE Wireless Communications*, *28*(6), 192-199. doi: 10.1109/MWC.001.2100102.

[6] Zhu, S., Ota, K., & Dong, M. (2021). Green AI for IIoT: Energy efficient intelligent edge computing for industrial internet of things. *IEEE Transactions on Green Communications and Networking*, *6*(1), 79-88. doi: 10.1109/TGCN.2021.3100622.

[7] Chohan, B. S., Xu, X., & Lu, Y. (2022). MES Dynamic interoperability for SMEs in the Factory of the Future perspective. *Procedia CIRP*, *107*, 1329-1335. https://doi.org/10.1016/j.procir.2022.05.153

[8] Dieguez, T., Malheiro, M. T., Leal, N., & Machado, J. (2025, June). Systematic Literature Review on Manufacturing Execution Systems in the Era of Industry 4.0: A Bibliometric Analysis. In *International Conference Innovation in Engineering* (pp. 298-310). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-94484-0_24

[9] Savaglio, C., Mazzei, P., & Fortino, G. (2024). Edge intelligence for industrial iot: Opportunities and limitations. *Procedia Computer Science*, *232*, 397-405. https://doi.org/10.1016/j.procs.2024.01.039

[10] Ma, J., Chen, H., Zhang, Y., Guo, H., Ren, Y., Mo, R., & Liu, L. (2020). A digital twin-driven production management system for production workshop. *The International Journal of Advanced Manufacturing Technology*, *110*(5), 1385-1397. https://doi.org/10.1007/s00170-020-05977-5

[11] Nguyen, V. H., Tran, Q. T., Besanger, Y., Jung, M., & Nguyen, T. L. (2022). Digital twin integrated power-hardware-in-the-loop for the assessment of distributed renewable energy resources. *Electrical Engineering*, *104*(2), 377-388. https://doi.org/10.1007/s00202-021-01246-0

[12] Knaeuper, A., & Rouse, W. B. (2012). A rule-based model of human problem-solving behavior in dynamic environments. *IEEE transactions on systems, man and cybernetics*, (6), 708-719. doi: 10.1109/TSMC.1985.6313454.

[13] Bharot, N., Verma, P., Soderi, M., & Breslin, J. G. (2024). DQ-DeepLearn: data quality driven deep learning approach for enhanced predictive maintenance in smart manufacturing. *Procedia Computer Science*, *232*, 574-583. https://doi.org/10.1016/j.procs.2024.01.057

[14] ElBaih, M. (2023). The role of privacy regulations in ai development (A Discussion of the Ways in Which Privacy Regulations Can Shape the Development of AI). *Available at SSRN 4589207*. https://dx.doi.org/10.2139/ssrn.4589207

[15] Danishvar, M., Danishvar, S., Katsou, E., Mansouri, S. A., & Mousavi, A. (2021). Energy-aware flowshop scheduling: A case for AI-driven sustainable manufacturing. *IEEe Access*, *9*, 141678-141692. doi: 10.1109/ACCESS.2021.3120126.

[16] Chitty-Venkata, K. T., & Somani, A. K. (2022). Neural architecture search survey: A hardware perspective. *ACM Computing Surveys*, *55*(4), 1-36. https://doi.org/10.1145/3524500

[17] Luong, N. H., Phan, Q. M., Vo, A., Pham, T. N., & Bui, D. T. (2024). Lightweight multi-objective evolutionary neural architecture search with low-cost proxy metrics. *Information Sciences*, *655*, 119856. https://doi.org/10.1016/j.ins.2023.119856

[18] Li, H., He, X., Wu, Y., Liu, G., Wang, H., Wen, X., & Li, L. (2026). Digital twin and AI-driven robotic embodied control system: a novel adaptive learning and decision optimization method. *Robotics and Computer-Integrated Manufacturing*, *98*, 103138. https://doi.org/10.1016/j.rcim.2025.103138

[19] Madi, A., Stan, O., Mayoue, A., Grivet-Sébert, A., Gouy-Pailler, C., & Sirdey, R. (2021, May). A secure federated learning framework using homomorphic encryption and verifiable computing. In *2021 Reconciling Data Analytics, Automation, Privacy and Security: A Big Data Challenge (RDAAPS)* (pp. 1-8). IEEE. doi: 10.1109/RDAAPS48126.2021.9452005.

[20] Lesi, V., Jakovljevic, Z., & Pajic, M. (2021). Security analysis for distributed IoT-based industrial automation. *IEEE Transactions on Automation Science and Engineering*, *19*(4), 3093-3108. doi: 10.1109/TASE.2021.3106335.

[21] Giovanni Battista Gaggero, Alessandro Armellin (2024). ICS-ADD -A Smart Industry Testbed Dataset for Cyber-Physical Security Monitoring Testing. IEEE Dataport. https://dx.doi.org/10.21227/4zht-tr07

[22] https://www.kaggle.com/datasets/paresh2047/uci-semcom

Blockchain Governance Framework and Assessment Tools from a Readiness Perspective

Nur Indrawati[1], Dana Indra Sensuse[2], Deden Sumirat Hidayat[3], Erisvaha Kiki Purwaningsih[4]
Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia[1, 2, 3, 4]
Ministry of Communications and Digital Affairs, Jakarta, Indonesia[1, 4]
National Research and Innovation Agency, Jakarta, Indonesia[3]

***Abstract*—Blockchain has been used in various sectors and use cases, but inadequate blockchain governance can cause failure in blockchain adoption. Previous research on blockchain governance has not been carried out with multiple perspectives and has not yet reached the point of developing governance assessment tools. This research aims to identify the factors that comprise blockchain governance, as well as identifying functional requirements of blockchain governance assessment tools. The methodology used in this research is literature review using PRISMA, thematic analysis, expert validation with fuzzy Delphi method. This research resulted in 25 factors and enhanced the STOPE framework by adding a collaboration dimension to develop a blockchain governance framework and 5 functional requirements for blockchain governance assessment tools. The 25 factors consist of 3 factors in strategy, 3 factors in technology, 4 factors in organization, 1 factor in the people, 4 factors in the environment, and 10 factors in the collaboration. This research identifies the factors that comprise blockchain governance and the functional requirements of blockchain governance assessment tools and provides a readiness-based blockchain governance initial framework, and functional requirements of blockchain governance assessment tools. This research also provides guidance for researchers, regulators, and practitioners in blockchain governance implementation and assessment.**

Keywords*—*Blockchain governance; blockchain governance framework; blockchain readiness; fuzzy Delphi method; governance assessment tools; STOPE framework

I. Introduction

Blockchain is an emerging technology that allows digital transactions without the role of intermediaries [1]. Blockchains that adopt the distributed ledger principle are, in certain cases, used because of their security or cryptographic features [2],[3] and privacy support [4]. However, there is quite a lot of research regarding the use of blockchain for cryptocurrency [5], conceptualizes blockchain as a trusted infrastructure to enhance community participation and strengthen local economies through cryptocurrencies [6], and analyzes the psychological factors influencing individuals' intention to hold or own Bitcoin [7].

Now the use of blockchain is increasingly widespread in various sectors and use cases. This use takes into account various blockchain characteristics, not only hashing or cryptography, but also immutability, decentralization, and transparency [8]. Blockchain is used in various use cases or sectors [13], such as non-fungible tokens (NFT) [9], access to Artificial Intelligence (AI) models [10], supply chain management [11], asset management [12], trading and privacy [13].

However, the current increase in the use of blockchain has not been matched by comprehensive research on its governance aspects [14],[15],[16],[17],[18]. Poor or inadequate governance in blockchain adoption or utilization is a factor that can cause blockchain adoption or utilization to fail [14]. Research on blockchain governance has been widely carried out using various perspectives. Research conducted by Tan, Mahula, & Crompvoets [14] proposes a blockchain governance conceptual framework in the context of the public sector, where blockchain governance should be understood as a socio-technical system rather than merely a technological implementation. Research by Laatikainen, Li, & Abrahamsson [15] conceptualize blockchain governance from a system perspective, in which blockchain is considered a dynamic and interconnected system, requiring holistic rather than partial analysis. Research by Beck, Müller-Bloch, & King [16] develops a foundational framework for governance in the blockchain economy and identifies two forms of governance, namely on-chain governance and off-chain governance. Research by Liu et al. [17] focuses on formulating principles of blockchain governance. Meanwhile, research by Koning & van der Linden [18] compares decentralized governance with corporate governance. However, the studies are still limited to the concept of governance, and the resulting framework is not equipped with properties for implementing governance based on the framework that has been researched [14],[15],[16],[17],[18]. Apart from that the framework built does not emphasize the readiness aspect or feasibility analysis stage [14],[15],[16],[17],[18], which is an important stage in developing a blockchain-based system [19].

Research on blockchain governance framework has been also carried out using various perspectives. Research conducted by Tan et al. [14] proposes a blockchain framework from the public management perspective. Research by Laatikainen et al. [15] develops blockchain governance model from a system-based perspective, which can serve as a reference framework for developing a blockchain system governance framework. Research by Beck et al. [16] develops blockchain governance from a corporate governance perspective. These studies have produced frameworks from different perspectives, but are less comprehensive, and have not been equipped with assessment properties or assessment tools for implementing governance following with the proposed framework [14],[15],[16].

The blockchain governance framework has been the focus of several studies, using various perspectives or theories. Research

by Pereira et al. [20] develops blockchain governance from the perspective of platform governance. Meanwhile, research by Laatikainen et al. [15] develops blockchain governance with a focus on an Information Technology governance perspective. Research by Beck et al. [16] develops blockchain governance from the perspective of Information Technology governance and corporate governance. Some studies view blockchain governance from other perspectives, such as game theory [45] and public management [14].

Based on the conditions described above, this research proposes an initial readiness-based multi-perspective and collaborative blockchain governance framework, which is equipped with the functional requirements of its assessment tools. This research aims to provide an initial blockchain governance framework and functional requirements for assessment tools. This research aims to provide guidance in readiness-based, multi-perspective, and collaborative blockchain governance so that it can prevent failure in blockchain adoption in various sectors or use cases. The research questions to be answered in this research are (1) What factors comprise blockchain governance framework? (2) What are the functional requirements of blockchain governance assessment tools?

This research contributes to developing a readiness-based, collaborative, and multi-perspective blockchain governance initial framework, equipped with the functional requirements of assessment tools, based on the initial framework developed. This research also enhances the STOPE framework [21],[22] with a collaboration dimension based on the characteristics [8] of blockchain technology which requires collaboration in its implementation.

This paper is organized as follows. Section II elaborates related work; Section III explains the research methodology; Section IV presents the research results and discussion; Section V closes the research with the conclusion and limitation of the study.

II. Related Work

A. Blockchain Governance Framework

The blockchain governance framework has been the focus of several studies, using various perspectives or theories. Research by Pereira et al. [20] develops blockchain governance from the perspective of platform governance. Meanwhile, research by Laatikainen et al. [15] develops blockchain governance with a focus on an Information Technology governance perspective. Research by Beck et al. [16] develops blockchain governance from the perspective of Information Technology governance and corporate governance. Some studies view blockchain governance from other perspectives, such as game theory [23] and public management [14]. A summary of literature review that represents these perspectives, key factors, strengths, limitations, and research gaps can be seen in Table I.

As summarized in Table I, existing blockchain governance remain fragmented, with each approach focusing on specific dimensions such as technology, incentives, or organizational control [14],[15],[16],[20],[23]. No prior studies have developed a fully integrated framework that simultaneously and systematically address strategic, technological, organizational, people, and environmental aspects [14],[15],[16],[20],[23]. This fragmentation indicates a significant research gap in the development of a comprehensive governance model capable of capturing the complexity of blockchain ecosystems. Therefore, this study proposes an enhanced STOPE-based framework to bridge this critical gap by providing a holistic and structured approach to blockchain governance.

TABLE I. Summarize of Literature Review on Blockchain Governance Framework

Perspective	Key Focus	Strengths	Limitations	Gap
Platform governance [20]	Coordination mechanisms	Explaining ecosystem coordination and roles	Coverage of factors.	Limited coverage of organizational and environmental factors.
IT governance [15],[16]	Alignment between IT structures and blockchain systems.	Relationship with organizational IT governance practices	Decentralization dynamics and external environment	Does not address multi-dimensional governance complexity
Corporate governance [16]	Accountability, control, and decision rights	Integrates managerial and control aspects	Technological adaptability and stakeholder dynamics	Partial integration; lacks environmental & people dimensions
Game theory [23]	Incentives (game theory)	Insights into specific governance aspects (incentives)	Fragmented	Not integrated governance aspects
Public management [14]	Regulation and public policy	Insights into specific governance aspects (regulation)	Highly domain-specific.	Not integrated governance aspects

B. Stope Framework and Collaboration

STOPE integrates the dimensions of strategy, technology, organization, people, and environment. The STOPE framework is a framework that is commonly used to measure readiness, for example, user readiness in implementing certain services or technologies [22]. The STOPE framework in the governance area can also be used as a view in assessments, such as in research to measure IT governance [21]. Using grades for factors evaluation will measure the readiness level for each factor. Table II describes the evaluation grade that can be used in readiness assessment [21],[22]. Evaluation of factors in STOPE is carried out individually and collectively [21],[22].

TABLE II. Evaluation Grades for Factors Evaluation

Scale	Evaluation Grade [21],[22]
1	Poor/low
2	Below average
3	Average
4	Above average
5	Good/high

In this research, the STOPE perspective is used by adding the collaboration dimension [8] to group factors identified by thematic analysis [24]. The addition of this dimension is related to the characteristics of blockchain which can be viewed not only from an Information Technology perspective but also from socio-technical aspects which include collaboration aspects [8]. A description of each dimension used in this research, in the context of its influence on blockchain governance, can be seen in Table III.

TABLE III. DEFINITION OF STOPE AND COLLABORATION DIMENSION

Dimension	Description and Representative Literature
Strategy	How the strategy aspect affects blockchain governance [25],[26],[27]
Technology	How the technology aspect affects blockchain governance [25],[27],[28]
Organization	How the organization aspect affects blockchain governance [25],[26],[27],[28]
People	How the people or personal aspect affects blockchain governance [29],[30]
Environment	How the environment aspect affects blockchain governance [25],[26],[28],[31]
Collaboration	How the collaboration aspect affects blockchain governance [25],[26],[27],[28],[30],[32]

The governance theory perspective [16],[18],[32] emphasizes that effective governance in multi-stakeholder environments relies on interaction processes, including joint decision-making, collective problem-solving, and consensus-building mechanisms. These processes cannot be fully represented within the existing STOPE [21],[22] dimensions, as they extend beyond individual domains and more accurately reflect the relational nature of governance through collaboration. Grouping collaboration-related factors within domains such as "people" or "organization" risks reducing collaboration to a localized attribute, thereby overlooking its systemic significance.

In the context of blockchain governance, the importance of collaboration is further reinforced by the absence or minimal presence of centralized control and the existence of trustless interactions. Mechanisms within blockchain systems, such as consensus protocols, decentralized decision-making, and community-based governance, inherently depend on collaborative processes [16],[18],[32]. Collaboration is therefore not merely a supporting factor, but a fundamental governance mechanism underpinning the functioning of the entire ecosystem [16],[18],[32].

Modeling collaboration as an enhancing dimension in this study explicitly captures the cross-cutting and integrative nature of stakeholder interactions, coordination processes, and interdependencies across STOPE [21],[22] dimensions. Enhancing the STOPE framework [21],[22] by introducing the collaboration dimension enables a more comprehensive and structurally coherent representation of blockchain governance, thereby enhancing the analytical depth and practical relevance of the proposed framework, particularly in evaluating multi-stakeholder governance systems.

C. Fuzzy Delphi Method

The Fuzzy Delphi method is a refinement of the Delphi method combined with fuzzy [33],[34]. The Fuzzy Delphi method reduces the possibility of ambiguity by consensus and reduces data retrieval time when compared to the Delphi method [33],[34]. The Fuzzy Delphi method can be used for forecasting, validation, and evaluation [33],[34].

D. Governance Assessment Tools

Governance assessment tools are usually used for environmental governance e.g. flood risk maps [35] and infrastructure [40], water governance [36], protected area [37],[38],[41], waste reduction [39], water, energy, and food [42]. One of the widely adopted instruments is the Governance Assessment Tool (GAT), which evaluates governance quality based on four key parameters: extent, coherence, flexibility, and intensity [40],[41],[42],[43]. Meanwhile, the components assessed, as can be seen in Table IV, include responsibilities and resources, actors and networks, strategies and instruments, levels and scales, problem perspectives and goal ambitions [35],[36],[37],[38],[39],[40],[41],[42]. This component can be considered in governance assessments in other areas or sectors.

TABLE IV. COMPONENT OF GOVERNANCE ASSESSMENT TOOLS

Component	Representative Literature
Responsibilities and resources	[35],[36],[37],[38],[39],[40],[41],[42]
Actors and networks	
Strategies and instruments	
Levels and scales	
Problem perspectives and goal ambitions	

III. METHODOLOGY

This section will discuss the materials and methods used in this research. This research employed mixed methods. This research has used the following methods: literature review, component and functional requirements analysis, thematic analysis [24], expert review with fuzzy Delphi questionnaire, and fuzzy Delphi analysis [33],[34].

A. Literature Review

The systematic literature review in this research was carried out in two stages, and both were carried out using the PRISMA method [44]. The steps for a systematic literature review in this research using the PRISMA method [44] include searching the database with defined criteria, determining inclusion criteria, study selection, and data synthesis. The first stage of the systematic literature review was carried out to identify previous research regarding governance assessment tools in general. The second stage of the systematic literature review was carried out to identify factors that influence blockchain governance. Apart from that, a literature review was also carried out to look for indicators of factors that influence blockchain governance.

1) Search strategy: In the first stage of the systematic literature review, the search was carried out on the Scopus database for journal articles published in all years. The terms or keywords used in the search are "blockchain governance". The

results of the systematic literature review in the first stage are influencing factors of blockchain governance.

The first stage of the systematic literature review was conducted using the Scopus database due to its extensive multidisciplinary coverage and rigorous indexing of peer-reviewed scholarly publications. Scopus is widely recognized as one of the largest curated abstract and citation databases, encompassing high-quality journals in information systems, computer science, governance, blockchain, and related socio-technical domains that are central to this study. Although other databases such as Web of Science, IEEE Xplore, and ACM Digital Library also contain relevant publications, Scopus was selected as the primary data source in the initial stage of the systematic literature review (SLR) due to its broad coverage and significant overlap with these databases, thereby ensuring adequate representation of high-quality literature. This approach is consistent with systematic review studies that rely on a single comprehensive database to maintain consistency and reduce duplication.

The search in the second stage of the systematic literature review was conducted across eight databases—Scopus, ScienceDirect, ACM Digital Library, Emerald Insight, ProQuest, IEEE Xplore, SpringerLink, Sage Journals, and Taylor & Francis—targeting journal articles published between 2019 and 2023. The terms or keywords used in the search are "governance assessment" AND "tool". The results of the systematic literature review in the second stage used in this research are the components and functionality requirements of governance assessment tools.

2) Eligibility criteria: In the first stage of the systematic literature review, the author established three inclusion criteria where the literature must be available in full text and written in English (IC4), must report research findings published in reputable journals (IC5), and must focus on and discuss governance assessment tools or blockchain governance assessment tools (IC6). In the second stage of the systematic literature review, the author applied three inclusion criteria, requiring that the literature be available in full text and written in English (IC1), report research findings published in reputable journals (IC2), and focus on the topic of blockchain governance (IC3). There is a total of six inclusion criteria (IC) in the two stages of the systematic literature review.

3) Study selection: Study selection is carried out at both the first and second stages of the systematic literature review. Study selection in this research includes the following steps: searching based on a keyword or search string in each online database, checking and removing duplicate records, selecting the title and abstract of identified articles according to the eligibility criteria, and including the articles that meet the inclusion criteria for full-text review. The study selection process is the same between the first stage and the second stage of systematic literature review in this research.

4) Data items and synthesis: Data collection was carried out manually using a spreadsheet. The information extracted in the first stage of the systematic literature review consists of design and development methods, elements (of readiness) measure, type of governance measure, assessment concept, and main features or main functionality. The information extracted in the second stage of the systematic literature review consists of influencing factors of blockchain governance. Authors then synthesize the data based on the STOPE framework [21],[22].

B. Functional Requirement Identification and Validation

Functional requirements identification [19] was carried out in this research to identify the functionality of the assessment tools based on the results of the first stage of the systematic literature review. Apart from being based on the research results, functional analysis also considers existing tools used in governance assessment practices [43]. After that, the results of the functional requirements identification are then reviewed by experts to ensure functional suitability (completeness, correctness, appropriateness) [45] of the functional requirements of the assessment tools.

C. Thematic Analysis

Thematic analysis [24] was carried out in this research to identify factors that influence blockchain governance based on the results of the literature review in the second stage. The stages in the thematic analysis are familiarization with data, initial code generation, theme search, review of themes, naming, and defining themes. The grouping of the initial stages of thematic analysis in this document is carried out based on the STOPE framework [21],[22],[24].

D. Expert Review with Fuzzy Delphi Method

Expert reviews are carried out on functional analysis, factors, and indicators. The criteria for selecting experts are having experience in the field of governance or policy formulation for at least five years and having a basic understanding of blockchain and information system development. An expert review of factors influencing blockchain governance was carried out based on the Fuzzy Delphi Method [33],[34].

Questionnaires whose readability has been validated are then distributed to experts. Semi-structured interviews were conducted to validate data from questionnaires filled in by experts. The results are used in Fuzzy Delphi Method analysis [33],[34].

A total of five experts participated in the Fuzzy Delphi process. Despite the relatively small panel size, the experts were purposively selected based on their domain expertise and professional experience in blockchain and governance. Their organizational backgrounds include both government and non-government sectors, as summarized in Table V.

Although the panel consisted of five experts, prior studies have demonstrated that fuzzy Delphi method can yield reliable results with small but carefully selected expert groups, particularly in specialized domains where expertise is more critical than sample size [33],[34]. The experts were selected based on the following criteria: 1) relevant expertise in blockchain governance, 2) professional experience, and 3) demonstrated contributions, such as publications, projects, or leadership roles in related fields. This selection process ensured that all participating experts possessed sufficient domain knowledge to provide informed and credible judgments.

TABLE V. EXPERTS' PROFILES

Expert ID	Organization	Position	Experience
Expert 1	Government	IT Policy Analyst	>15 years
Expert 2	Non-government	IT Governance Specialist	>10 years
Expert 3	Government	IT Policy Analyst	>15 years
Expert 4	Non-government	Vice President of Technology	>5 years
Expert 5	Non-government	IT Governance Specialist	>25 years

In this research, the fuzzy Delphi method was used because this method can accommodate various opinions and uncertainties [33],[34]. The fuzzy Delphi method is suitable for research where the availability of appropriate information does not exist, or the information is difficult to obtain [33],[34]. This method can be used for forecasting and filtering [33],[34].

The Fuzzy Delphi Method is an expert consensus-based method with the important concept being the Triangular Fuzzy Number or TFN (n1, n2, n3) [33],[34]. Each TFN (n1, n2, n3) from the expert responses was averaged and produced values m1, m2, and m3 [33],[34]. Acceptable item requirements are a threshold (d) of 0.2 or less, an expert agreement of 75% or greater, and a defuzzification value of 0.5 or more [33],[34]. The conversion of Likert scores into fuzzy numbers and linguistic variables [33],[34] can be seen in Table VI. Table VI shows the Likert scale used is a 7-point Likert scale [33],[34].

TABLE VI. CONVERSION BETWEEN SEVEN-POINT LIKERT SCALE AND FUZZY NUMBER

Likert Scale	Linguistic Variable	Fuzzy Number		
		n1	*n2*	*n3*
1	Extremely Strongly Disagree	0.0	0.0	0.1
2	Strongly Disagree	0.0	0.1	0.3
3	Disagree	0.1	0.3	0.5
4	Unsure	0.3	0.5	0.7
5	Agree	0.5	0.7	0.9
6	Strongly Agree	0.7	0.9	1.0
7	Extremely Strongly Agree	0.9	1.0	1.0

The process of calculating d (threshold) is carried out using the following formula:

$$d(\underline{m}, \underline{n}) = \sqrt{\frac{1}{3}(m_1 - n_1)^2 + (m_2 - n_2)^2 + (m_3 - n_3)^2} \quad (1)$$

Meanwhile, the defuzzification value calculation is carried out using the following formula:

$$A = \frac{1}{3} * (m1 + m2 + m3) \quad (2)$$

If the d value is 0.2 or less, and the A value is 0.5 or more, the item or factor is retained.

E. Framework and Assessment Tools Development and Validation

Based on the results of factor validation using the Fuzzy Delphi Method [33],[34], the results of identification and validation of indicators, analysis, and validation of functional requirements from assessment tools, then an initial blockchain governance framework and assessment tools were created. Next, expert validation is carried out on the proposed blockchain governance framework, to ensure the perceive ease of use, usefulness, extent, coherence and flexibility [35],[36],[37],[38],[39],[40],[41],[42],[46].

This study does not adopt the STOPE framework as a measurement model or rating scale; rather, it is employed as a structural framework to organize governance dimensions. The assessment is conducted using ISO/IEC-based criteria, thereby ensuring a standardized and rigorous evaluation. This integration enables a comprehensive alignment between multidimensional governance aspects and established evaluation standards.

IV. RESULTS AND DISCUSSION

This section will discuss the results of this research and its discussion. The findings and their implications are also discussed in this section. The limitations of this research are highlighted and recommendations for future research are given in this section.

A. Influencing Factors of Blockchain Governance

The first step in identifying factors that influence blockchain governance is to carry out a systematic literature review using the PRISMA method [44], on the Scopus database for journals published all year. The stages in the selection of previous studies consist of identification, screening, and inclusion [44]. Following the removal of duplicates and the screening of titles, abstracts, and keywords, the remaining articles were further selected according to the inclusion criteria, i.e. must be available in full text and written in English (IC4), must report research findings published in reputable journals (IC5), and must focus on and discuss governance assessment tools or blockchain governance assessment tools (IC6). Details of the article selection process can be seen in Fig. 1.

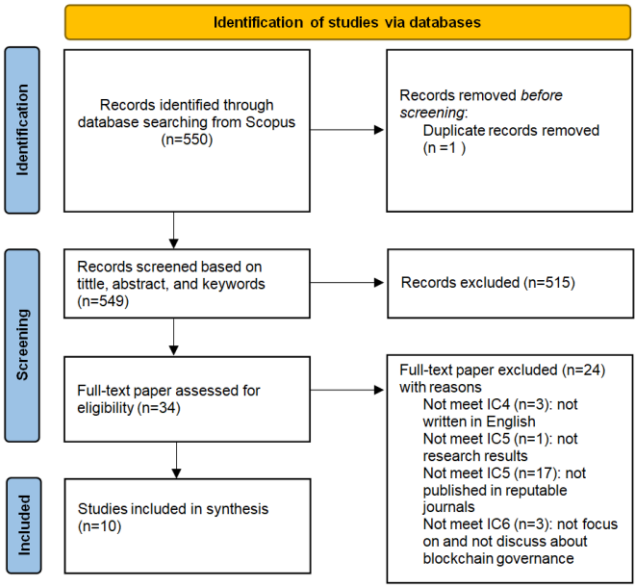


Fig. 1. Flow diagram for search results on the first stage of systematic literature review.

Fig.1 describes the search and filtering process for article selection. A total of 550 articles were obtained by entering the keyword "blockchain governance" in the Scopus database. After eliminating duplicates and filtering based on title, abstract, and keywords, 34 articles were obtained which were filtered based on eligibility criteria. Screening by eligibility criteria resulted in 10 articles being included for synthesis. Synthesis results from the 10 articles and analysis with thematic analysis [46] can be seen in Table VII for the strategy dimension, Table VIII for the technology dimension, Table IX for the organization dimension, Table X for the people dimension, Table XI for the environment dimension, and Table XII for the collaboration dimension.

Table VII describes the influencing factors in the strategy dimension. Five factors were identified within the strategy dimension, including incentives, centralized risk, flexibility of governance, allocation, and the enaction of decision rights [25],[26],[27],[50]. These five strategy factors can influence blockchain governance [25],[26],[27],[50].

TABLE VII. INFLUENCING FACTOR OF STRATEGY DIMENSION

Factor	Code	Representative Literature
Motivational factors (incentive)	S1	[25]
Nature of how decision rights are allocated and enacted	S2	[26]
Centralization risk	S3	[27]
Flexibility of governance	S4	[27],[50]
Misalignment of incentives	S5	[27]

Table VIII presents the influencing factors in the technology dimension. Within the technology dimension, three factors were identified, including formation and contextual elements of the blockchain, such as background information (e.g., licensing), resolve of conflict between competing version of ledger, and inter-proposal dependencies [25],[26],[28],[50]. These three technological factors have the potential to impact blockchain governance [25],[26],[28],[50].

TABLE VIII. INFLUENCING FACTOR OF TECHNOLOGY DIMENSION

Factor	Code	Representative Literature
Formation and context	T1	[25]
Conflicts resolving	T2	[28],[50]
Inter-proposal dependencies	T3	[26]

TABLE IX. INFLUENCING FACTOR OF ORGANIZATION DIMENSION

Factor	Code	Representative Literature
Roles, responsibility, accountability	O1	[25]
Security of DAO	O2	[28]
Problem at hand	O3	[26]
Corporate governance	O4	[25][28]

Table IX describes the influencing factors in the organization dimension. Four factors were identified within the organizational dimension, including roles, responsibilities, and accountability [25], security of DAO [28], and the specific problem at hand [26], such as demand management, data management, system architecture design and development, as well as corporate governance [25][27]. These four organization factors can influence blockchain governance [25],[26],[27],[28].

Table X describes the influencing factors in the people dimension. Two factors were identified within the people dimension, comprising dilemmas of sharing knowledge and sharing information resources, and vulnerability of the combination of private interests [29],[30]. These two people factors can influence blockchain governance [29],[30].

TABLE X. INFLUENCING FACTOR OF PEOPLE DIMENSION

Factor	Code	Representative Literature
Social dilemmas (sharing knowledge and information resources)	P1	[30]
Vulnerability of the combination of private interests	P2	[29]

Table XI describes the influencing factors in the environment dimension. Four factors were identified within the environmental dimension, including legal aspect, organizational theories, public or private sector role, and the nature of information exchanged [25],[26],[28],[31]. These four environmental factors can influence blockchain governance [25],[26],[28],[31].

TABLE XI. INFLUENCING FACTOR OF ENVIRONMENT DIMENSION

Factor	Code	Representative Literature
The nature of the information exchanged	E1	[28]
Legal compliance	E2	[31]
Organizational theories	E3	[26]
Role of public or private sector	E4	[25]

Table XII describes the influencing factors in the collaboration dimension. Eleven factors were identified within the collaboration dimension, including membership or way participation, membership management for available roles, and decision-making processes [25]. These eleven collaboration factors can influence blockchain governance [25],[28],[30],[32],[47],[50].

TABLE XII. INFLUENCING FACTOR OF COLLABORATION DIMENSION

Factor	Code	Representative Literature
Membership	C1	[25]
Communication	C2	[25],[50]
Decision making	C3	[25]
Coordination mechanism within a group of users	C4	[28]
Decision problem	C5	[26]
Consensus of stakeholders	C6	[27]
Respons in problem cases	C7	[27]
Exercise of voting rights	C8	[47]
Community governance	C9	[30]
Private property	C10	[30]
Democrate form of governance	C11	[32]

The factors that have been generated from the systematic literature review and thematic analysis are then analyzed based on the Fuzzy Delphi Method [33],[34]. The factor is retained if the threshold d = 0.2 or less, and the defuzzification value A = 0.5 or greater than that, and the expert's agreement is 75% or greater than that [33],[34]. The analysis results based on the Fuzzy Delphi Method [33],[34] can be seen in Table XIII for the strategy dimension, Table XIV for the technology dimension, Table XV for the organization dimension, Table XVI for the people dimension, Table XVII for the environment dimension, and Table XVIII for the collaboration dimension. In total, there were 29 factors proposed. Two factors of which were accepted although they did not meet the threshold criteria (d), due to the expert agreement exceeding 75%. Two factors of which were rejected because the threshold value (d) is greater than 0.2 and the percentage of expert agreement is less than 75%. A total of 27 factors were accepted.

Table XIII describes the calculation results using the fuzzy Delphi method for factors in the strategy dimension. There are five factors proposed in the strategy dimension. Two of these factors, specifically S3 (centralization risk) and S5 (misalignment of incentives), were retained despite their threshold values (d) exceeding 0.2, as expert consensus surpassed 75%.

As described in Table XIII, the accepted factors on strategy dimension in this research are in line with previous research conducted by Pelt et al. [25], Ziolkowski et al. [26], and Dursun et al. [27]. However, in contrast with research conducted by Dursun et al. [27], the centralization risk factor and misalignment of incentives in this research are rejected based on expert consensus in the fuzzy Delphi method. This rejection may be influenced by the work experience and organizational background of the expert involved in filling out the questionnaire in this research and may be due to the scope of the research conducted by Dursun et al. [27] is limited to on-chain governance, so there is a difference in risk.

TABLE XIII. RESULTS ON THE STRATEGY DIMENSION

Factor Code	Threshold value (d)	Percentage of Expert Agreement	Fuzzy Score (A)	Expert Consensus
S1	0.073	100%	0.927	Accept
S2	0.073	100%	0.927	Accept
S3	0.312	80%	0.760	Reject
S4	0.118	100%	0.873	Accept
S5	0.235	80%	0.673	Reject

TABLE XIV. RESULTS ON THE TECHNOLOGY DIMENSION

Factor Code	Threshold value (d)	Percentage of Expert Agreement	Fuzzy Score (A)	Expert Consensus
T1	0.168	100%	0.840	Accept
T2	0.168	100%	0.840	Accept
T3	0.188	100%	0.807	Accept

Table XIV describes the calculation results using the fuzzy Delphi method for factors in the technology dimension. There are three factors proposed in the technology dimension. These three factors are accepted based on expert consensus, with a threshold value (d) of less than 0.2 and a fuzzy score (A) of greater than 0.8. The accepted factors on the technology dimension in this research are in line with previous research conducted by Pelt et al. [25], Dursun et al. [27], and Maggiolino et al. [28].

Table XV describes the results of calculations using the fuzzy Delphi method for factors in the organizational dimension. There are four factors proposed in the organizational dimension. These four factors are accepted based on expert consensus, with a threshold value (d) of less than 0.2 and a fuzzy score (A) of greater than 0.8, where the minimum value of fuzzy score (A) for an item or factor to be accepted is 0.5. The accepted factors on organization dimension in this research are in line with previous research conducted by Pelt et al. [25], Ziolkowski et al. [26], Dursun et al. [27], and Maggiolino et al. [28].

TABLE XV. RESULTS ON THE ORGANIZATION DIMENSION

Factor Code	Threshold value (d)	Percentage of Expert Agreement	Fuzzy Score (A)	Expert Consensus
O1	0.132	100%	0.893	Accept
O2	0.188	100%	0.860	Accept
O3	0.132	100%	0.893	Accept
O4	0.132	100%	0.893	Accept

Table XVI describes the results of calculations using the fuzzy Delphi method for factors in the people dimension. There are two factors proposed in the people dimension. One factor, P1 (dilemmas of sharing knowledge and sharing information resources), was rejected because its threshold value (d) exceeded 0.2 and the level of expert agreement was below 75%. This rejection may be influenced by the organizational background and organizational culture of each expert, which may cause differences in views regarding aspects of sharing knowledge and sharing information resources as mentioned in research conducted by Murtazashvili et al. [30]. Factor P2 is accepted and aligns with research conducted by Reijers et al. [29].

The lack of consensus on P1 indicates that social dilemmas in the context of knowledge sharing and information resources are not perceived as a dominant factor in blockchain governance by the expert panel. In contrast, P2 (vulnerability of the combination of private interests) was identified as the factor with the highest level of agreement, highlighting the vulnerability arising from the interaction of diverse private interests within decentralized systems. This finding reflects a fundamental characteristic of blockchain ecosystems, where independent actors with heterogeneous objectives operate without a centralized authority, thereby increasing the potential for interest conflicts and opportunistic behavior.

From a theoretical perspective, these results can be understood through socio-technical perspectives [14] and collaborative governance theory [16],[51],[52], which emphasize that the primary challenges in decentralized systems are not limited to technical coordination, but also include the management of competing interests. Within this framework, P2 (vulnerability of the combination of private interests) represents critical dynamics such as agency problems, collective action

challenges, and incentive misalignments, which are often key sources of vulnerability in community-based governance systems. Therefore, although only one factor was retained, P2 (vulnerability of the combination of private interests) still reflects a broad and fundamental construct in explaining the human dimension of blockchain governance.

TABLE XVI. RESULTS ON THE PEOPLE DIMENSION

Factor Code	Threshold value (d)	Percentage of Expert Agreement	Fuzzy Score (A)	Expert Consensus
P1	0.255	40%	0.573	Reject
P2	0.168	100%	0.840	Accept

Table XVII describes the calculation results using the fuzzy Delphi method for factors in the environmental dimension. There are four factors proposed in the environmental dimension. These four factors are accepted based on expert consensus, with a threshold value (d) of less than 0.2 and a fuzzy score (A) of greater than 0.8. The four factors in the environmental dimension that were accepted in this study are in line with previous research conducted by Pelt et al. [25], Ziolkowski et al. [26], Maggiolino et al. [28], and De Filippi et al. [31].

TABLE XVII. RESULTS ON THE ENVIRONMENT DIMENSION

Factor Code	Threshold value (d)	Percentage of Expert Agreement	Fuzzy Score (A)	Expert Consensus
E1	0.168	100%	0.840	Accept
E2	0.168	100%	0.840	Accept
E3	0.168	80%	0.840	Accept
E4	0.168	100%	0.840	Accept

Table XVIII describes the calculation results using the fuzzy Delphi method for factors in the collaboration dimension. There are eleven factors proposed in the collaboration dimension. One factor, C11 (a democratic form of governance), was rejected its threshold value (d) exceeded 0.2 and the level of expert agreement was below 75%. The democratic form of governance reject factor may be due to the case in the research conducted by Thompson [10] taking a case study of digital currency, where the use of digital currency has not been regulated and has not been formally implemented in the organization or environment of origin of the experts. The ten factors proposed in this collaboration dimension were accepted based on expert consensus using the fuzzy Delphi method analysis, and this is in line with research that has been conducted previously by Pelt et al. [25], Ziolkowski et al. [26], Dursun et al. [27], and Maggiolino et al. [28], Venugopalan et al. [47], and Murtazashvili et al. [30].

TABLE XVIII. RESULTS ON THE COLLABORATION DIMENSION

Factor Code	Threshold value (d)	Percentage of Expert Agreement	Fuzzy Score (A)	Expert Consensus
C1	0.188	100%	0.807	Accept
C2	0.188	100%	0.807	Accept
C3	0.168	100%	0.840	Accept
C4	0.168	100%	0.840	Accept
C5	0.132	100%	0.893	Accept
C6	0.145	100%	0.820	Accept
C7	0.132	100%	0.893	Accept
C8	0.188	100%	0.807	Accept
C9	0.155	100%	0.787	Accept
C10	0.132	100%	0.893	Accept
C11	0.362	40%	0.567	Reject

B. Readiness-Based Bloccckhain Governance Framework

Based on the results of a systematic literature review, thematic analysis, and fuzzy Delphi method analysis, an initial readiness-based blockchain governance framework was built. This proposed framework was built from a readiness perspective, with enhanced STOPE framework. The initial framework comprises six domains derived from the STOPE framework perspective [21],[22] and blockchain characteristics [8], including the strategy, technology, organization, people, environment, and collaboration domains. The proposed readiness-based blockchain governance framework can be seen in Fig. 2 below.

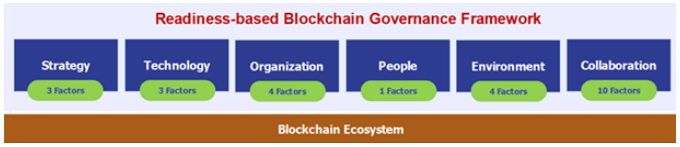


Fig. 2. Proposed readiness-based blockchain governance framework.

Fig. 2 describes the initial readiness-based blockchain governance framework resulting from a systematic literature review using the PRISMA method [44], thematic analysis [8],[21],[22],[24], and analysis using the fuzzy Delphi method [33],[34], and expert validation. The developed framework encompasses six dimensions: strategy, technology, organization, people, environment, and collaboration. In each framework, factors are the result of expert consensus based on the fuzzy Delphi method [33],[34]. In the strategy dimension, there are 3 factors, in the technology dimension there are 3 factors, in the organization dimension there are 4 factors, in the people dimension there is 1 factor, and in the environment dimension there are 10 factors.

For each factor in each dimension, indicators are identified for assessing blockchain governance readiness. In this research, the factors whose indicators have been identified are DAO security (O2). The selection of factors is based on expert recommendations from interview results. Identification of indicators for the security of DAO factor (O2) was carried out using a literature review. Indicators for the O2 security of the DAO factor, which have been identified from ISO/IEC 27002:2022 [48] and have been validated by experts, are policies for information security, information security roles and responsibilities, segregation of duties, management responsibilities, and information security in project management.

C. Functionality of Assessment Tools

Fig. 3 describes the first stage of the systematic literature review process using the PRISMA method [44]. This systematic

literature review was carried out to find functional requirements for blockchain governance assessment tools. The search was carried out in ten databases for journal articles published between 2019 and 2023.

Searches were conducted on the databases Scopus, ScienceDirect, ACM, Digital Library, Emerald Insight, ProQuest, IEEE Xplore, SpringerLink, Sage Journals, and Taylor & Francis. At the identification stage, 555 articles were obtained and 10 of them were duplicates so that 545 articles were processed at the screening stage. The screening stage referring to inclusion criteria IC1, IC2, and IC3 resulted in 24 studies included in the synthesis.

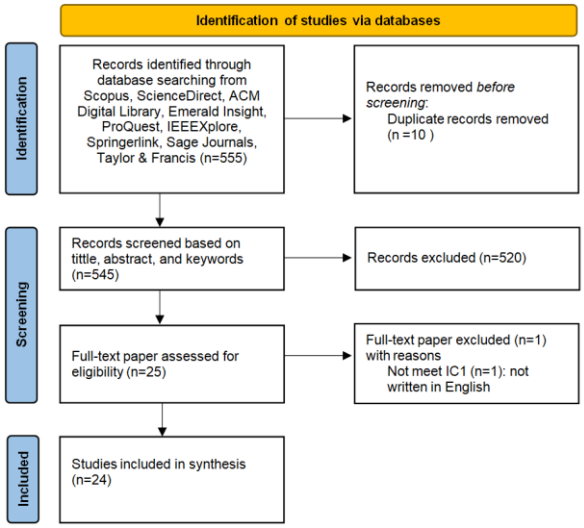


Fig. 3. Flow diagram for search results on the second stage of systematic literature review.

The results of the second-stage systematic literature review revealed a set of functional features, including strategy adaptation, assessment, good practices, community building, and supporting information. The functional results obtained are then combined with the results of the systematic literature review in the second stage and then illustrated in Fig. 4 to make it easier to understand and develop.

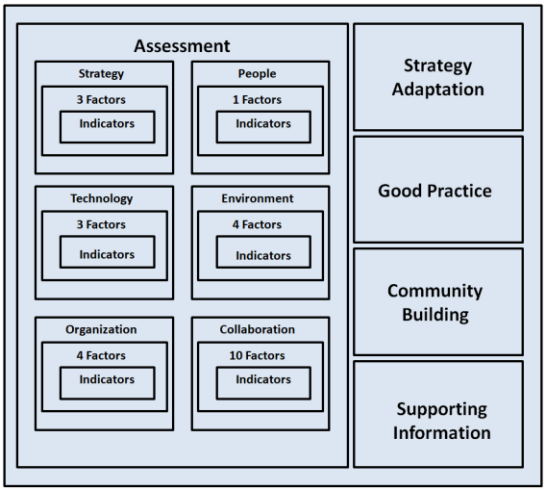


Fig. 4. List of functional requirements of blockchain governance assessment tools.

Fig. 4 presents the set of functional requirements for blockchain governance assessment tools, comprising five key functionalities: assessment, strategy adaptation, good practices, community building, and supporting information. The assessment tools encompass six dimensions: strategy, technology, organization, people, environment, and collaboration. In each dimension, there are factors and indicators for assessment. The "strategy adaptation" functionality is a functionality to inform the strategy adapted in the designed assessment tools. Meanwhile, the "good practice" functionality is functionality to inform about blockchain governance good practice. The "community building" functionality is related to the development of blockchain communities, both technical and non-technical aspects. This list of functional requirements is then subjected to expert validation.

In this study, ISO/IEC 38503:2022 Information technology—Governance of IT—Assessment of the governance of IT [53] was selected as the reference model and rating scale for the assessment instrument based on its relevance in evaluating governance within complex socio-technical systems. Although ISO/IEC 38503:2022 [53] was originally developed for information technology (IT) governance, it provides an assessment approach and structure that can be applied to blockchain environments, which inherently operate as decentralized and multi-stakeholder digital infrastructures. Blockchain governance extends beyond purely technical mechanisms to encompass organizational, behavioral, and environmental aspects, thereby requiring a comprehensive evaluation framework capable of capturing these interdependencies.

ISO/IEC 38503:2022 Information technology—Governance of IT—Assessment of the governance of IT [53] provides a robust assessment structure through its emphasis on governance tasks and practices, evidence of success, and the achievement of beneficial outcomes. This structure enables the operationalization of abstract governance concepts into measurable components, thereby addressing a common limitation in existing blockchain governance studies, which tend to remain conceptual and lack clear evaluative mechanisms. Furthermore, this international standard reflects a global consensus among key stakeholder groups, namely government, industry, academia, and end users, thereby enhancing the credibility and generalizability of the assessment model and rating scale. The adoption of ISO/IEC 38503:2022 Information technology—Governance of IT—Assessment of the governance of IT in this study ensures that the proposed measurement instrument is not only theoretically grounded but also aligned with globally recognized governance practices. This alignment strengthens the methodological rigor of the study and supports the development of a systematic and multi-dimensional assessment of blockchain governance across the STOPE and collaboration domains.

Table XIX presents the measurement instrument for blockchain governance assessment tools. Each factor within the domains of strategy, technology, organization, people, environment (STOPE), and collaboration is assessed based on the ISO/IEC 38503:2022 Information technology—Governance of IT—Assessment of the governance of IT rating scale and maturity model [53]. The assessment is conducted across three

key governance aspects, i.e. the implementation of governance tasks and practices as the fundamental aspect of governance, evidence of success, and the achievement of beneficial outcomes [53]. A five-point rating scale is used in the assessment, where 1 indicates unknown, 2 not applied, 3 somewhat applied, 4 largely applied, and 5 fully applied [53]. This five-point rating scale adopted from ISO/IEC 38503:2022 Information technology—Governance of IT—Assessment of the governance of IT [53].

TABLE XIX. MEASUREMENT INSTRUMENT FOR BLOCKCHAIN GOVERNANCE ASSESSMENT TOOLS

Factor	Assessment	Rating				
		1	2	3	4	5
Dimension: Strategy						
Motivational factors (incentive)	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Nature of how decision rights are allocated and enacted	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Flexibility of governance	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Dimension: Technology						
Formation and context	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Conflicts resolving	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Inter-proposal dependencies	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Dimension: Organization						
Roles, responsibility, accountability	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Security of DAO	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Problem at hand	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Corporate governance	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Dimension: People						
	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Dimension: Organization						
The nature of the information exchanged	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Legal compliance	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Organizational theories	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Role of public or private sector	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Dimension: Collaboration						
Membership	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Communication	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Decision making	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Coordination mechanism within a group of users	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Decision problem	Are governance tasks and practices being applied?					
	Is there evidence of success?					

Factor	Assessment	Rating				
		1	2	3	4	5
	Are beneficial outcomes being achieved?					
Consensus of stakeholders	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Respons in problem cases	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Exercise of voting rights	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Community governance	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Private property	Are governance tasks and practices being applied?					
	Is there evidence of success?					
	Are beneficial outcomes being achieved?					
Rating Scale 1 Unknown 2 Not applied 3 Somewhat applied 4 Largely applied 5 Fully applied						

D. Comparison with Prior Work

In comparison with prior studies on blockchain governance [14],[15],[16],[20],[23], this research adopts a more comprehensive set of perspectives, encompassing platform governance, information technology governance, corporate governance, game theory, public management, and collaboration. As depicted in Table XX, research conducted by Pereira et al. [20] perceives blockchain governance from the perspective of platform governance and collaboration, while research conducted by Laatikainen et al. [15] perceives blockchain governance from an IT governance perspective. Table XX also describes the research conducted by Beck et al. [16] perceives blockchain governance from the perspective of Information Technology governance and corporate governance, while research conducted by Lee et al. [23] perceives blockchain from a game theory perspective, and research conducted by Tan et al. [14] perceive blockchain governance from a public management and collaboration perspective. Despite the diversity of perspectives presented in prior studies, they remain largely fragmented and domain-specific, lacking a unified framework that integrates these viewpoints into a comprehensive model of blockchain governance.

Compared with governance frameworks for other innovative and emerging technologies, such as artificial intelligence governance examined by de Almeida et al. [49], this study aligns with the emphasis on incorporating risk considerations in governance development. This research is also in line with research conducted by de Almeida et al. [49] in considering the characteristics of technology to be governed or regulated. When viewed from a regulatory perspective, this research is also in line with research conducted by de Almeida et al. [49] for its usefulness in highlighting regulatory points.

TABLE XX. COMPARISON OF PERSPECTIVE USED IN THE STUDY

Perspective	[20]	[15]	[16]	[23]	[14]	This study
Platform governance	√	x	x	x	x	√
IT governance	x	√	√	x	x	√
Corporate governance	x	x	√	x	x	√
Game theory	x	x	x	√	x	√
Public management	x	x	x	x	√	√
Collaboration	√	x	x	x	√	√

E. Implications

1) Theoretical implications: This research makes four theoretical contributions. First, this research develops a readiness-based, collaboration, and multiperspective blockchain governance framework. To the best of our knowledge, there has been no research that builds blockchain governance based on readiness, collaboration, and multiperspective [14],[15],[16],[20],[23]. Therefore, with this research, we fill the research gap that is important to ensure success in blockchain adoption [14]. Second, we enhanced the stope point of view, which is based on the stope framework [21],[22], based on the results of our literature analysis. This enhancement is done by adding a collaboration dimension [8]. This addition was based on technology characteristics, risk considerations, and previous research [8],[49]. Third, our research results in a list of functional requirements for blockchain governance assessment tools. The list of functional requirements was obtained from the analysis we carried out on previous research of governance assessment [35],[36],[37],[38],[39],[40],[41],[42]. The results of our research can be used as a reference in developing information system development theory in the future, for example, by considering technological characteristics, while also considering risks. Fourth, we provide a future research agenda in the blockchain governance subfield. Future research agendas can use more varied perspectives than existing research. Future research agendas can also use more varied empirical data for stakeholder engagement.

2) Practical implications: This research has two practical implications. First, the blockchain governance initial framework and list of functional requirements can be a reference for practitioners who will adopt blockchain or develop blockchain governance assessment tools. Second, the blockchain governance initial framework which contains factors, can be a reference for regulator in determining the main points of blockchain regulation or policy.

V. CONCLUSIONS

This research identifies and discusses factors comprising blockchain governance and functional requirements of blockchain governance assessment tools. This research provides a readiness-based blockchain governance initial framework, built from 25 factors grouped based on the enhanced STOPE framework point of view. This research also provides functional requirements for blockchain governance assessment. This research also provides valuable information that can be a reference for researchers in conducting their research, regulators to highlight the main point of blockchain governance or regulation, and practitioners in blockchain governance framework and assessment in line with governance and risk compliance concept.

A. Limitations of the Study

In this study, the identification of factors that influence blockchain governance through a systematic literature review was limited to English language literature. This study uses only a single database (Scopus) for systematic literature review in identifying factors that influence blockchain governance. This study also does not include readiness level evaluation and maturity level.

B. Future Work

Future research needs to be conducted regarding the readiness and maturity of blockchain governance. Future research should focus on assessing readiness levels and extending this study using broader methodologies and data sources, including the operationalization of indicators for the factors identified. Identification of indicators can be done not only from previous research but can come from international standards or gray literature such as industry reports or news. In future research, more stakeholders can be engaged, with more representative backgrounds and work experience. In addition, to overcome the shortcomings of the fuzzy Delphi method which may have an impact on the quality of the data obtained, it can be carried out in more than one iteration. The findings of this study represent the collective judgment of the selected experts at a specific point in time and may not fully capture the continuously evolving perspectives within the field. Future research is encouraged to involve a larger and more diverse expert panel or to adopt alternative consensus techniques in order to further strengthen the validity of the findings. Further investigation into blockchain governance assessment tools remains necessary. In addition, future studies should address maturity aspects, particularly the identification of relevant indicators and the determination of maturity levels.

ACKNOWLEDGMENT

In writing this journal, the ChatGPT application has been used by the author to help with the translation of part of the text from Indonesian to English. This research was funded by the Directorate General of Digital Infrastructure, Ministry of Communications and Digital Affairs of the Republic of Indonesia.

REFERENCES

[1] A. Gurzhii, A. K. M. N. Islam, A. K. M. B. Haque, and V. Marella, "Blockchain enabled digital transformation: A systematic literature review," IEEE Access, vol. 10, pp. 79584–79605, 2022, doi: 10.1109/ACCESS.2022.3194004.

[2] M. S. Ahammad, M. Maliha, N. E. Nila, et al., "An innovative blockchain framework for strengthening security and efficiency in banking," Sci. Rep., vol. 15, p. 39029, 2025, doi: 10.1038/s41598-025-25457-8.

[3] S. Almarri and A. Aljughaiman, "Blockchain technology for IoT security and trust: A comprehensive SLR," Sustainability, vol. 16, no. 23, p. 10177, 2024, doi: 10.3390/su162310177.

[4] M. Zirui and G. Bin, "A privacy-preserved and user self-governance blockchain-based framework to combat COVID-19 depression in social media," IEEE Access, vol. 11, pp. 35255–35280, 2023, doi: 10.1109/ACCESS.2023.3264598.

[5] F. Liu, H. Y. Fan, and J. Y. Qi, "Blockchain technology, cryptocurrency: Entropy-based perspective," Entropy, vol. 24, no. 4, p. 557, 2022, doi: 10.3390/e24040557.

[6] E. H. Diniz, A. K. Cernev, D. A. Rodrigues, and F. Daneluzzi, "Solidarity cryptocurrencies as digital community platforms," Inf. Technol. Dev., vol. 27, no. 3, pp. 524–538, 2021, doi: 10.1080/02681102.2020.1827365.

[7] K. Rzayev, A. Sakkas, and A. Urquhart, "An adoption model of cryptocurrencies," Eur. J. Oper. Res., vol. 323, no. 1, pp. 253–266, 2025, doi: 10.1016/j.ejor.2024.11.024.

[8] V. Ali, A. A. Norman, and S. R. bin Azzuhri, "Characteristics of blockchain and its relationship with trust," IEEE Access, vol. 11, pp. 15364–15374, 2023, doi: 10.1109/ACCESS.2023.3243700.

[9] N. Alnuaimi, A. Almemari, M. Madine, K. Salah, H. al Breiki, and R. Jayaraman, "NFT certificates and proof of delivery for fine jewelry and gemstones," IEEE Access, vol. 10, pp. 101263–101275, 2022, doi: 10.1109/ACCESS.2022.3208698.

[10] A. Battah, M. Madine, I. Yaqoob, K. Salah, H. R. Hasan, and R. Jayaraman, "Blockchain and NFTs for trusted ownership, trading, and access of AI models," IEEE Access, vol. 10, pp. 112230–112249, 2022, doi: 10.1109/ACCESS.2022.3215660.

[11] A. Mohammed, V. Potdar, M. Quaddus, and W. Hui, "Blockchain adoption in food supply chains: A systematic literature review on enablers, benefits, and barriers," in IEEE Access, vol. 11, pp. 14236–14255, 2023, doi: 10.1109/ACCESS.2023.3236666.

[12] V. T. Truong, L. Le, and D. Niyato, "Blockchain meets metaverse and digital asset management: A comprehensive survey," IEEE Access, vol. 11, pp. 26258–26288, 2023, doi: 10.1109/ACCESS.2023.3257029.

[13] F. A. Sunny, P. Hajek, M. Munk, M. Z. Abedin, M. S. Satu, M. I. A. Efat, and M. J. Islam, "A systematic review of blockchain applications," in IEEE Access, vol. 10, pp. 59155–59177, 2022, doi: 10.1109/ACCESS.2022.3179690.

[14] E. Tan, S. Mahula, and J. Crompvoets, "Blockchain governance in the public sector: A conceptual framework for public management," Gov. Inf. Q., vol. 39, no. 1, 2022, doi: 10.1016/j.giq.2021.101625.

[15] G. Laatikainen, M. Li, and P. Abrahamsson, "A system-based view of blockchain governance," Inf. Softw. Technol., vol. 157, 2023, doi: 10.1016/j.infsof.2023.107149.

[16] R. Beck, C. Müller-Bloch, and J. L. King, "Governance in the blockchain economy: A framework and research agenda," J. Assoc. Inf. Syst., vol. 19, no. 10, pp. 1020–1034, 2018, doi: 10.17705/1jais.00518.

[17] Y. Liu, Q. Lu, G. Yu, H. Y. Paik, and L. Zhu, "Defining blockchain governance principles: A comprehensive framework," Inf. Syst., vol. 109, 2022, doi: 10.1016/j.is.2022.102090.

[18] H. Koning and T. van der Linden, "Decentralized governance inspired by corporate governance?" Proc. IEEE 24th Conf. Bus. Inform. (CBI 2022), vol. 2, pp. 104–110, 2022, doi: 10.1109/CBI54897.2022.10055.

[19] J. L. Whitten and L. D. Bentley, Systems Analysis and Design Methods, 7th ed., 2007.

[20] J. Pereira, M. M. Tavalaei, and H. Ozalp, "Blockchain-based platforms: Decentralized infrastructures and its boundary conditions," Technol. Forecast. Soc. Change, vol. 146, pp. 94–102, 2019, doi: 10.1016/j.techfore.2019.04.030.

[21] H. Bin-Abbas and S. H. Bakry, "Assessment of IT governance in organizations: A simple integrated approach," Comput. Hum. Behav., vol. 32, pp. 261–267, 2014, doi: 10.1016/j.chb.2013.12.019.

[22] H. Choi, M. J. Park, J. J. Rho, and H. Zo, "Rethinking the assessment of e-government implementation in developing countries from the perspective of the design–reality gap: Applications in the Indonesian e-procurement system," Telecommun. Policy, vol. 40, no. 7, pp. 644–660, 2016, doi: 10.1016/j.telpol.2016.03.002.

[23] B. E. Lee, D. J. Moroz, and D. C. Parkes, "The political economy of blockchain governance," SSRN Scholarly Paper ID 3537314, 2020, doi: 10.2139/ssrn.3537314.

[24] V. Braun and V. Clarke, Thematic Analysis: A Practical Guide. SAGE Publications Ltd, 2022.

[25] R. van Pelt, S. Jansen, D. Baars, and S. Overbeek, "Defining blockchain governance: A framework for analysis and comparison," Inf. Syst. Manag., vol. 38, no. 1, pp. 21–41, 2021, doi: 10.1080/10580530.2020.1720046.

[26] R. Ziolkowski, G. Miscione, and G. Schwabe, "Decision problems in blockchain governance: Old wine in new bottles or walking in someone else's shoes?" J. Manag. Inf. Syst., vol. 37, no. 2, pp. 316–348, 2020, doi: 10.1080/07421222.2020.1759974.

[27] T. Dursun and B. B. Üstündağ, "A novel framework for policy based on-chain governance of blockchain networks," Inf. Process. Manag., vol. 58, no. 4, 2021, doi: 10.1016/j.ipm.2021.102556.

[28] M. Maggiolino and L. Zoboli, "Blockchain governance: The missing piece in the competition puzzle," Comput. Law Secur. Rev., vol. 43, 2021, doi: 10.1016/j.clsr.2021.105609.

[29] W. Reijers, I. Wuisman, M. Mannan, P. de Filippi, C. Wray, V. Rae-Looi, A. Cubillos Vélez, and L. Orgad, "Now the code runs itself: On-chain and off-chain governance of blockchain technologies," Topoi, vol. 40, no. 4, pp. 821–831, 2021, doi: 10.1007/s11245-018-9626-5.

[30] I. Murtazashvili, J. B. Murtazashvili, M. B. H. Weiss, and M. J. Madison, "Blockchain networks as knowledge commons," Int. J. Commons, vol. 16, no. 1, pp. 108–119, 2022, doi: 10.5334/ijc.1146.

[31] P. de Filippi, M. Mannan, and W. Reijers, "The alegality of blockchain technology," Policy Soc., vol. 41, no. 3, pp. 358–372, 2022, doi: 10.1093/polsoc/puac006.

[32] G. Thompson, "Democracy, citizenship, and corporate governance reform: How to deal with the internationalization of corporate activity," Thesis Eleven, vol. 167, no. 1, pp. 42–57, 2021, doi: 10.1177/07255136211058481.

[33] P. L. Chang, C. W. Hsu, and P. C. Chang, "Fuzzy Delphi method for evaluating hydrogen production technologies," Int. J. Hydrogen Energy, vol. 36, no. 21, pp. 14172–14179, 2011, doi: 10.1016/j.ijhydene.2011.05.045.

[34] H. F. Sulaiman, R. Ismail, H. Mohd Yusoff, N. Anuar, M. R. Mohd Jamil, and F. Daud, "Validation of occupational zoonotic disease questionnaire using fuzzy Delphi method," J. Agromed., vol. 25, no. 2, pp. 166–172, 2020, doi: 10.1080/1059924X.2019.1666763.

[35] C. C. Flores and J. Crompvoets, "Assessing the governance context support for creating a pluvial flood risk map with climate change scenarios: The Flemish subnational case," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 7, p. 460, 2020, doi: 10.3390/ijgi9070460.

[36] M. Lordkipanidze, H. Bressers, and K. Lulofs, "Comparative assessment of water governance in protected areas," *Water*, vol. 12, no. 3, p. 740, 2020, doi: 10.3390/w12030740.

[37] C. Casiano, V. Vikolainen, and J. Crompvoets, "Governance assessment of a blue-green infrastructure project in a small size city in Belgium: The potential of Herentals for a leapfrog to water sensitive," Cities, vol. 117, p. 103331, 2021, doi: 10.1016/j.cities.2021.103331.

[38] M. Lordkipanidze, H. Bressers, and K. Lulofs, "Governance assessment of a protected area: The case of the Alde Feanen National Park," J. Environ. Plan. Manag., pp. 1–24, 2019, doi: 10.1080/09640568.2018.1441014.

[39] M. D. Latanna, B. Gunawan, M. L. Franco-García, and H. Bressers, "Governance assessment of community-based waste reduction program in Makassar," Sustainability, vol. 15, no. 19, 2023, doi: 10.3390/su151914371.

[40] C. C. Flores, J. Crompvoets, M. E. Ibarraran Viniegra, and M. Farrelly, "Governance assessment of the flood's infrastructure policy in San Pedro Cholula, Mexico: Potential for a leapfrog to water sensitive," *Sustainability*, 2019.

[41] S. J. Mirnezami, C. De Boer, and A. Bagheri, "Groundwater governance and implementing the conservation policy: The case study of Rafsanjan Plain in Iran," Environ. Dev. Sustain., vol. 22, no. 8, pp. 8183–8210, 2020, doi: 10.1007/s10668-019-00488-0.

[42] T. W. John, J. Sušnik, S. Masia, and G. Jewitt, "Towards realization of nexus-doing at the grassroots level: Water-energy-food governance assessment in the Songwe River Basin (Tanzania and Malawi)," *Environ. Sci. Policy*, vol. 150, p. 103596, 2023, doi: 10.1016/j.envsci.2023.103596.

[43] A webtool for municipalities and citizens," [Online]. Available: www.catch-tool.com.

[44] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," The BMJ, vol. 372, 2021, doi: 10.1136/bmj.n71.

[45] Systems and software engineering - Systems and software — Quality requirements and evaluation (SQuaRE) — Guide to SQuaRE, ISO/IEC 25000:2014, 2014.

[46] R. Marais, S. Grobbelaar, and I. de Kock, "Healthcare technology transfer in sub-Saharan Africa: Conceptual framework evaluation," IEEE Trans. Eng. Manag., 2021, doi: 10.1109/TEM.2021.3109445.

[47] S. Venugopalan, I. Stancikova, and I. Homoliak, "Always on voting: A framework for repetitive voting on the blockchain," IEEE Trans. Emerg. Top. Comput., 2023, doi: 10.1109/TETC.2023.3315748.

[48] Information security, cybersecurity and privacy protection — Information security controls, ISO/IEC 27002:2022, 2022.

[49] P. G. R. de Almeida, C. D. dos Santos, and J. S. Farias, "Artificial intelligence regulation: A framework for governance," Ethics Inf. Technol., vol. 23, no. 3, pp. 505–525, 2021, doi: 10.1007/s10676-021-09593-z.

[50] J. Parkin, "The senatorial governance of Bitcoin: making (de)centralized money," Economy Soc., vol. 48, no. 4, pp. 463–487, 2019, doi: 10.1080/03085147.2019.1678262.

[51] B. A. Jnr, "Toward a collaborative governance model for distributed ledger technology adoption in organizations," Environ. Syst. Decis., vol. 42, no. 2, pp. 276–294, 2022, doi: 10.1016/j.envsci.2021.12.003.

[52] F. Beyers, J. Leventon, and H. Heinrichs, "Collaborative governance or state regulation? Endless efforts but little capacity for sustainability transformation of the German textile sector," Entrep. Sustain. Issues, 2022, doi: 10.1007/s13668-021-00396-9.

[53] ISO, "ISO/IEC 38503:2022 Information technology—Governance of IT—Assessment of the governance of IT," 2022.

The Untapped Potential of Extended Reality for Indigenous Medicinal Knowledge: A Review of Cross-Disciplinary XR Applications

Giampearo Anak Peter[1], Suriati Khartini Jali[2], Jane Labadin[3],
Abby Lian Hendrick[4], Ally Dian Hendrick[5], Nurul Farizah Ridzuan[6]
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan , Malaysia[1, 3, 4, 5, 6]
Institute for Tourism Research and Innovation, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia[2]

***Abstract*—Many sectors are quickly moving toward XR for education and professional training. These technologies, which include Virtual, Augmented, and Mixed Reality, are not being applied equally across different types of knowledge preservation. This study aims to review how effective these technologies are and whether they have overlooked the preservation of Indigenous Medicinal Knowledge. Following the PRISMA 2020 guidelines, relevant literature from 2019 to 2026 was gathered from major academic databases, including IEEE Xplore, Scopus, and ScienceDirect. After removing duplicates and non-English papers, 23 out of 39 publications met the selection criteria. The findings show that while XR is a proven technology in many sectors, it remains underused for documenting the medicinal knowledge of communities such as the Kayan and Kenyah. This lack of progress is not caused by technical limitations, but by internal organizational barriers and weak project planning. The evidence, therefore, points to a procedural problem rather than a technological one.**

***Keywords*—Extended Reality; Indigenous Medicinal Knowledge; Virtual Reality; Augmented Reality; Mixed Reality; gamification; cultural heritage preservation; Task-Technology Fit; PRISMA**

I. Introduction

Indigenous medicinal knowledge is considered an important foundation for cultural history and community well-being. It includes traditional healing methods and plant knowledge that have been passed down orally across many generations [1]. In this system, medicinal uses are combined with the ecological and cultural values that shape Indigenous worldviews. One example of this can be seen in the Kenyah and Kayan communities of Sarawak, Malaysia, where plant identification and preparation are part of traditional healing practices that have long supported community health. However, this continuity is now under threat.

Traditional ways of passing down this knowledge have been disrupted by modernization, putting these practices at risk of being lost [1]. As younger generations adopt digital lifestyles and attend formal schools that are disconnected from their ancestral roots, these oral traditions face the danger of disappearing permanently.

Extended Reality (XR) is a technology that combines Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), and it can be proposed as a way to document and preserve this intangible cultural heritage. XR is an umbrella term that covers AR, which can be accessed through smartphones, tablets, or wearable devices such as the Microsoft HoloLens 2. AR layers images and text onto the user's real physical environment. VR, on the other hand, covers the user's entire field of vision through a headset and places them in a fully virtual space. Mixed Reality sits between AR, VR, and the real world [2].

XR technologies have been shown to create more immersive and effective learning experiences [2], [5]. They have been linked to better learning motivation, interest, creativity, and academic performance, while also being a more time-efficient option compared to traditional methods. XR is already being used as a training tool in medicine, the military, and sports education.

A notable gap exists: although XR is a well-established tool across many high-stakes fields such as chemistry, aviation, mental health, and medical training, it has rarely been applied to preserving Indigenous Medicinal Knowledge. In medicine, XR is used for neurosurgery and dental training [2], [7], [8], [9]. In chemistry education, it supports 3D molecular visualization [21]. Aviation uses it for flight simulation [5], and mental health practitioners apply it in cognitive rehabilitation and exposure therapy [13], [18]. The fact that XR is technically mature in these areas but largely absent from cultural preservation efforts suggests that the main obstacles are not technological, but procedural.

This review addresses the following research questions: 1) How is XR currently used for training in advanced professional fields such as medicine and aviation? 2) What are the main barriers to adopting XR, and how can inclusive design help address them? 3) How can XR technologies be adapted to preserve Indigenous Medicinal Knowledge in community settings?

II. Research Methodology

This systematic literature review was carried out using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 framework [3], which ensures that the process is transparent, reproducible, and methodologically sound. The review comprised four stages: identification, screening, eligibility, and inclusion.

A. Database Selection and Search Strategy

Three major databases were selected for their coverage of interdisciplinary research in computer science, cultural studies, and educational technology: IEEE Xplore, Scopus, and ScienceDirect. They were chosen because they provide access to peer-reviewed studies and include research on XR and gamification in education and cultural preservation. These three databases were considered sufficient for this review because they jointly index the majority of peer-reviewed venues publishing XR research, including IEEE VR, IEEE BigComp, MetroXRAINE, JMIR, Frontiers, and MDPI journals that appeared in the final included set. A preliminary coverage check confirmed that extending the search to additional databases (e.g., ACM Digital Library, Web of Science) returned substantial overlap with records already indexed in Scopus and IEEE Xplore. The review was restricted to English-language publications to ensure consistency in screening and appraisal; this restriction is acknowledged as a limitation in Section IV.E.

B. Search Strategy

The following Boolean search string was applied across all three databases:

"extended reality" AND ((("indigenous" OR "tradition*" OR "heritage") AND ("medicin*")) OR "ethnobotan*" OR "ethnopharma*")

A time restriction was applied to focus on studies published from 2019 to 2026. Only peer-reviewed publications written in English were included. After searching the three selected databases, 39 records were identified for evaluation.

C. Inclusion and Exclusion Criteria

Studies were included if they met the following criteria: 1) published from 2019 to 2026; 2) written in English and published in a peer-reviewed venue; 3) used XR technologies such as VR, AR, or MR in gamified or interactive applications; 4) focused on the preservation of Indigenous, cultural, or medicinal knowledge; and 5) used mobile or mobile-compatible XR applications that support accessibility and community participation.

Exclusion criteria comprised: 1) non-English or non-peer-reviewed works; 2) lack of relevance to XR in educational or cultural contexts; 3) non-interactive XR applications without engagement elements.

D. Screening and Selection Process

Initially, 39 unique records were identified: 25 from Scopus, 10 from IEEE Xplore, and 4 from ScienceDirect. Within the Scopus collection, 3 records were removed for not being written in English, 2 were identified as duplicates, and 11 did not meet the inclusion requirements, leaving 9 Scopus papers. IEEE Xplore (n=10) and ScienceDirect (n=4) records all met the inclusion criteria. After full-text analysis, 23 papers were selected: IEEE Xplore (n=10), Scopus (n=9), and ScienceDirect (n=4). The PRISMA flow diagram is shown in Fig. 1.

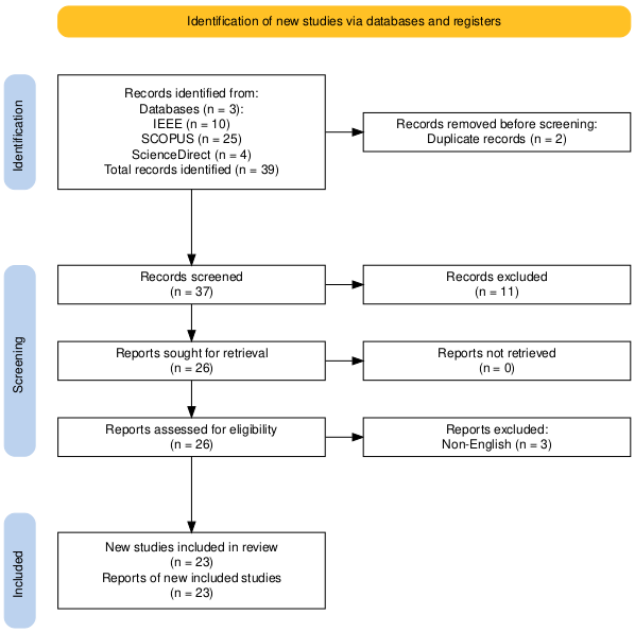


Fig. 1. PRISMA flow diagram

E. Data Extraction and Thematic Coding

After final selection, data extraction was carried out to gather key information from each of the 23 included studies. Each study was coded across several fields: 1) author(s) and publication year; 2) XR modality (AR, VR, MR, or hybrid); 3) presence of gamification elements; 4) domain of application, such as heritage, education, accessibility; 5) methodology used; 6) key outcomes; and 7) reported challenges or limitations.

F. Analytical Framework

This review uses the Task-Technology Fit (TTF) framework to assess how well XR technologies match the needs of knowledge preservation. TTF argues that a technology's effectiveness depends on how closely its features match the demands of the task [24]. The framework was originally developed to explain individual performance in information systems [4] and has since been applied to technology adoption in educational settings [24]. In knowledge preservation, the "task" includes cultural transmission, environmental learning, and community participation, while the "technology" refers to the different XR modalities and their capabilities.

To operationalize the TTF framework, this review explicitly maps four core task characteristics against the technology features offered by XR systems reported in the 23 included studies. The task characteristics derived from the domain of Indigenous Medicinal Knowledge (IMK) preservation are: (T1) spatial and sensory knowledge representation (e.g., plant identification, preparation gestures); (T2) narrative and oral transmission; (T3) community-based, in-field accessibility; and (T4) cross-generational engagement.

The corresponding technology features observed across the included studies are: (F1) immersive 3D visualization and spatial interaction [5], [17], [21]; (F2) scenario-based and simulation modalities [5], [19]; (F3) mobile and low-cost form factors [1], [6], [14]; and (F4) gamified feedback loops such as rewards, exploration, and progression [1], [5], [21]. Table V in Section IV-B presents the explicit task-to-feature mapping.

G. Quality Assessment

To strengthen methodological rigor, a post-hoc quality appraisal was applied to all 23 included studies after final selection. A five-item checklist adapted from common critical-appraisal tools was used, covering: (Q1) clarity of research aim and objectives; (Q2) appropriateness of the methodological approach; (Q3) transparency of data collection or study procedure; (Q4) validity and traceability of findings; and (Q5) relevance to XR and to knowledge-preservation or training contexts. Each item was scored as 1 (met), 0.5 (partially met), or 0 (not met). A study was retained in the synthesis if it scored at least 3.0 out of 5.0. All 23 studies met this threshold, which is consistent with their selection from peer-reviewed indexed venues (IEEE Xplore, Scopus, and ScienceDirect). This appraisal does not replace a full risk-of-bias assessment, which is typically reserved for interventional studies; the current review is descriptive and cross-disciplinary, and the checklist therefore focuses on methodological transparency and relevance rather than internal bias.

III. Results and Analysis

A. Overview of Included Studies

From the initial 39 records retrieved, the following were excluded from Scopus during screening: non-English language (n=3), duplicates (n=2), and papers that did not meet the inclusion criteria (n=11). All records from IEEE Xplore records (n=10) and ScienceDirect records (n=4) met the inclusion criteria. The final set included 23 peer-reviewed studies published from 2019 to 2026: IEEE Xplore (n=10), Scopus (n=9), and ScienceDirect (n=4). The included studies spanned clinical, educational, and assistive domains, covering the healthcare metaverse and next-generation care ecosystems [10], image-guided pain-procedure navigation [12], mixed-reality pre-surgical planning for epilepsy [15], digital-twin applications for pediatric assessment [20], adaptive-optics corrections in XR visualization [22], and VR-based ecological memory assessment [23].

B. XR Modalities and Applications

The 23 included studies show a range of XR applications across educational and cultural contexts. Augmented Reality (AR) and Mixed Reality (MR) were the most used modalities [1], [2], [5]. The studies spanned multiple disciplines, including educational technology, medical education, cultural heritage preservation, and accessibility frameworks [8], [9], [14]. Fig. 2 presents the distribution of XR modalities across application domains for the 23 included studies, and Table I summarizes the XR modalities and their primary applications.

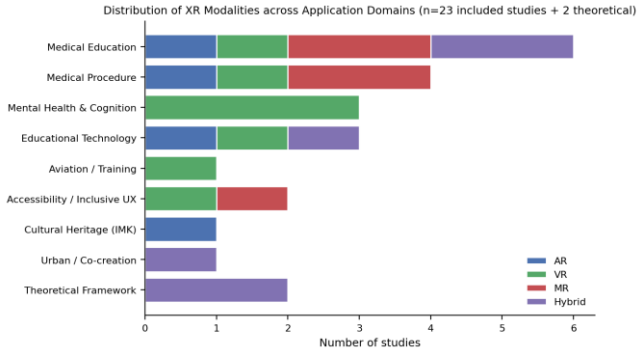


Fig. 2. Distribution of XR modalities (AR, VR, MR, Hybrid) across application domains in the 23 reviewed studies plus theoretical-framework references.

TABLE I. XR Modalities and Applications Across Included Studies

XR Modality	Primary Applications	Freq.	Key Characteristics	References
Augmented Reality (AR)	Educational enhancement; Cultural heritage visualization	High	Smartphone/tablet compatible; Real-world overlay; High accessibility	[1],[2],[5]
Mixed Reality (MR)	Medical training; Educational contexts	High	Immersive while maintaining spatial awareness; Wearable compatible	[5],[8]
Virtual Reality (VR)	Specialized training; Spatial memory assessment	Moderate	Full immersion; High cognitive engagement; Requires dedicated hardware	[5],[17],[18]
Hybrid (AR+VR+MR)	Complex medical and educational applications	Moderate	Combines AR/VR/MR capabilities; Flexible deployment	[1],[5],[9]

C. XR for Educational Enhancement

Several studies showed that XR technologies improve learning outcomes in educational settings. XR offers spatial understanding, greater peripheral awareness, and three-dimensional graphics, allowing for a level of realism that traditional classroom settings cannot provide [5], [17]. It has also been shown to strengthen cognitive ability, improve attention, and support neuroplasticity [5]. XR environments further encourage experimental and active learning [5], [21].

D. Gamification in XR Environments

Gamified XR applications were a strong category in the literature, with a focus on learning engagement, motivation, and retention. Game-based XR has been linked to higher scores in interest and enjoyment, perceived competence, and perceived usefulness [1]. Gamification elements such as progression, rewards, and exploration were found to be effective in encouraging curiosity and discovery-based learning, especially among younger audiences [1], [5], [21]. Table II presents the gamification mechanics identified across the reviewed studies.

TABLE II. GAMIFICATION MECHANICS IN XR APPLICATIONS

Mechanic	Description	Learning Impact	Cultural Fit	References
Interactive task completion	Users complete tasks to unlock narratives	High – promotes discovery	Excellent – aligns with quest traditions	[1],[5]
Achievement Systems	Rewards for mastering traditional practices	High – increases motivation	Good – recognizes mastery	[1],[5],[21]
Peer-based interaction and knowledge sharing	Peer-based knowledge sharing	High – promotes collaborative learning	Excellent – honors oral traditions	[1],[5],[21]
Immersive experiential learning	Exploration-driven experiences in educational contexts	High – immersive engagement	Excellent – connects to cultural narratives	[1],[5],[21]
Progression Systems	Level-based knowledge mastery	Moderate–High – clear goals	Good – structured learning path	[1],[5]

A cultural fit assessment represents the authors' analysis of potential alignment with Indigenous knowledge preservation needs.

E. Barriers to XR Adoption

Although XR applications have advanced technologically, their use in community settings remains limited due to technical, cultural, and ethical constraints. Common technical barriers include hardware costs, limited mobile compatibility, and poor infrastructure in rural areas [5], [2], [14]. From a TTF perspective, these barriers reflect a "mismatch", where technology's characteristics do not align with task requirements of remote, low-bandwidth environments [5], [4]. Other barriers include a lack of teacher training, few available educational experiences, limited research, and insufficient institutional support [2], [8], [9]. Table III summarizes the barriers identified.

TABLE III. BARRIERS TO XR ADOPTION IN COMMUNITY CONTEXTS

Barrier Category	Specific Barriers	Impact	TTF Analysis	References
Technical	Hardware costs; Limited mobile compatibility; Inadequate rural infrastructure	High	Technology mismatch: high-end requirements vs. low infrastructure	[5],[2],[14]
Infrastructure	High-speed internet; Stable power; Low bandwidth	High	Task–Technology mismatch: infrastructure unavailable	[4],[5]
Implementation	Lack of teacher training; Few educational experiences; No institutional support	Moderate–High	Organizational readiness issues	[2],[8],[9]
Economic	Initial hardware costs; Licensing fees; Maintenance expenses	High	Budget constraints for institutions	[2],[8],[9]

F. Opportunities for XR Integration

Despite the existing barriers, the review found emerging opportunities for XR in knowledge preservation. The potential of mobile-based XR platforms is supported by general XR education literature [6], and community-based testing has also been demonstrated [1]. Mobile-compatible XR platforms appear to be the most realistic option for community settings [1], [6]. Existing research also highlights the value of gamified and interactive approaches for passing down knowledge [1], [5], [21]. Table IV outlines the identified opportunities and their implementation strategies.

TABLE IV. OPPORTUNITIES FOR XR INTEGRATION IN KNOWLEDGE PRESERVATION

Opportunity	Implementation Strategy	Expected Outcome	Scalability	References
Mobile-Based XR	Deploy AR through smartphone/tablet apps	Accessible to remote communities	High – smartphone penetration	[1],[6]
Participatory Co-Creation	Co-design with community members and diverse age groups	Culturally authentic content and shared ownership	Moderate – requires engagement	[1]
Gamified Storytelling	Embed quests, achievements, collaborative gameplay	High engagement and knowledge retention	Moderate – effectiveness demonstrated in educational contexts	[1],[5],[21]

G. Usability and User Experience

Evaluations using the System Usability Scale (SUS) in existing literature show that XR applications generally achieve moderate to high usability scores [1]. Data from the Intrinsic Motivation Inventory (IMI) subscales indicated that users view XR tools as both effective and useful for learning, with particularly high scores in perceived competence and value [1]. A clear preference for AR over VR in certain environments also suggests that AR-based platforms may be more effective due to their ease of use and wider accessibility [1].

H. Accessibility and Inclusive Design

The literature also identified the importance of inclusive user experience design for diverse populations. Older adults and those with varying levels of digital literacy often find complex interfaces difficult to use, which points to the need for inclusive design approaches [14]. The digital divide continues to leave certain populations behind [2], [14]. Introducing XR in educational and community settings may help reduce this gap in access and adoption [2], [14].

IV. Discussion and Limitations

A. Theoretical Frameworks

The findings connect to several theoretical frameworks that link learning, technology, and knowledge preservation. XR technologies support experiential and active learning theory by allowing users to build understanding through immersive exploration [5], [17], [21]. Gamification elements also align with the intrinsic motivation indicators measured through the IMI [1]. Central to this discussion is the Task-Technology Fit (TTF) model, which looks at whether the capabilities of gamified mobile XR match the environmental and pedagogical tasks involved in knowledge preservation [4], [5].

B. Task-Technology Fit Analysis

The review also points to a clear mismatch in current research, where high-end XR hardware often requires infrastructure that is not available in rural or resource-limited settings [4], [5]. Task characteristics require that knowledge preservation supports mobile, in-field documentation within community environments [1], [6]. The task is not only technical but social, requiring a medium that supports oral, narrative, and practice-based transmission of knowledge [1]. Mobile XR platforms provide the portability and accessibility necessary to bridge the infrastructure gap [1], [6]. By utilizing mobile-compatible platforms, the objective of accessible community-based knowledge transmission is supported [1], [6].

TABLE V. Task-Technology Fit Mapping for XR in Indigenous Medicinal Knowledge Preservation

Task characteristic (IMK preservation)	Required capability	XR feature observed in reviewed studies	Supporting refs
T1. Spatial/sensory knowledge (plant ID, preparation gestures)	3D visualization, spatial interaction, gesture capture	Immersive VR/MR environments; stereoscopic XR video; acupoint annotation pipelines	[5], [11], [17], [21]
T2. Narrative/oral transmission	Scenario-based storytelling, audio-visual layering	AR overlays on real environments; simulation-based interprofessional scenarios	[2], [7], [19]
T3. Community-based, in-field accessibility	Portable, low-cost, low-bandwidth deployment	Mobile-compatible XR; low-cost VR for rehabilitation; inclusive UX for non-expert users	[1], [6], [14], [18]
T4. Cross-generational engagement	Motivation, retention, gamified feedback	Gamified XR with rewards and exploration; educational XR with measured motivation gains	[1], [5], [21]

C. Addressing Research Questions Through Findings

Table VI presents a synthesis of the research questions and the evidence-based answers derived from the 23 reviewed studies.

D. Practical Implications

The practical implications drawn from the synthesis focus on co-creation [16], accessibility, and ethical management of cultural content—areas essential for the sustainable preservation of knowledge [1], [6]. Table VII presents strategic recommendations for community-based XR implementation.

TABLE VI. Research Questions and Evidence-Based Answers

Research Question	Key Findings	Evidence	References
Current state of XR adoption for cultural heritage and education?	XR is becoming more common in educational institutions, though widespread adoption is still in its early stages. AR is generally preferred due to its accessibility through smartphones. Studies show that XR produces moderate to significant improvements in educational outcomes, including motivation, competence, and engagement.	Across the reviewed studies, educational settings consistently included engagement and interactivity features. Gamification was identified as a key design element in applications focused on cultural and educational purposes.	[1],[2],[5],[6],[8],[9],[14],[17]
What are the main barriers to adopting XR, and how can inclusive design overcome them?	Common barriers include hardware costs, limited mobile compatibility, inadequate rural infrastructure, lack of teacher training, and lack of institutional support. Inclusive design through mobile-based XR, participatory co-creation, and low-cost platforms can overcome these barriers.	TTF framework reveals technology-task mismatch in remote, low-bandwidth environments. Mobile XR platforms provide portability and accessibility to bridge the infrastructure gap.	[2],[4],[5],[8],[9],[14]
How can XR preserve and transmit knowledge in community contexts?	Mobile-based XR combined with gamification is optimal. Participatory co-design ensures cultural authenticity. Gamified storytelling bridges generational gaps.	Gamified mobile XR enhances knowledge retention and engagement. TTF framework explains adoption barriers and implementation solutions.	[1],[5],[6],[21]

TABLE VII. STRATEGIC RECOMMENDATIONS FOR COMMUNITY-BASED XR IMPLEMENTATION

Focus Area	Recommended Action	Intended Outcome	References
Community Participation	Co-design sessions with community members and diverse age groups	Ensure cultural accuracy and shared ownership of XR content	[1],[6]
Mobile XR Platform Selection	Select mobile-compatible XR platforms based on community technical capacity, budget, and ownership needs	Expand access in rural areas; ensure sustainability and community control over digital heritage	[1],[6]
Gamification Elements	Embed story-driven quests, rewards, and interactive feedback	Increase engagement and knowledge retention among youth	[1],[5],[21]
Ethical Data Governance	General XR literature recommends ethical and inclusive adoption	Safeguard knowledge ownership and promote trust in digital preservation	[6]

E. Limitations

While this systematic literature review followed PRISMA 2020 guidelines, several limitations must be acknowledged. The review was limited to three academic databases (IEEE Xplore, Scopus, ScienceDirect) and to English-language publications, which may have excluded regional, local, and Indigenous repository studies. The search string centered on the phrase "extended reality" combined with indigenous-, heritage-, and medicinal-related terms; it did not include the abbreviation "XR" as a separate keyword, nor narrower medicinal-domain terms such as "pharmacognosy", "healing", "herbal", "botany", or "plants". This conservative keyword set was a deliberate methodological choice to minimize the inclusion of studies unrelated to XR-for-knowledge-preservation, but it may have reduced recall. Future systematic reviews should expand the keyword set to include these terms, apply field-specific Boolean filters, and extend coverage to additional databases (e.g., ACM Digital Library, Web of Science) as well as regional and grey-literature sources. Because direct empirical research on Indigenous Medicinal Knowledge (IMK) preservation is scarce, the synthesis had to draw on broader educational and cultural heritage XR literature. All findings are strictly based on the 23 reviewed papers to maintain rigor and traceability. Table VIII provides a structured overview of these limitations.

TABLE VIII. LIMITATIONS OF THE SYSTEMATIC LITERATURE REVIEW

Limitation Category	Specific Limitation	Impact on Findings	Mitigation Strategy
Database & Language	Restricted to IEEE, Scopus, ScienceDirect; English-only	May exclude regional/local and Indigenous repository studies	Include grey literature, Indigenous repos, non-English publications in future reviews
Temporal	Review spans 2019–2026; rapid XR evolution	Earlier studies may reflect superseded hardware	Pedagogical/design insights remain relevant despite hardware evolution
Content Scope	Scarcity of direct IMK preservation research	Synthesis relied on broader XR/education literature	Future research should focus specifically on IMK preservation applications
Community Participation	Lack of explicit community participation in many studies	Limited culturally grounded XR frameworks discussed	Recommend participatory research designs with Indigenous communities
Methodological Scope	Review based on 23 papers from a systematic search only	All claims grounded strictly in reviewed papers	Ensures rigor and traceability; prevents unsupported claims
Research Gap	Limited direct IMK-specific XR empirical evidence	Recommendations extrapolated from general education/ cultural XR	Future empirical studies needed in community-based IMK contexts

F. Cultural, Ethical, and Epistemological Considerations

Although XR applications have advanced technically, their use for Indigenous Medicinal Knowledge (IMK) raises cultural, ethical, and epistemological concerns that must be explicitly addressed. Indigenous knowledge systems are often relational, place-based, and embedded in oral and ritual practice; digitizing such knowledge through XR risks decontextualizing it from the community protocols that govern who may hold, share, and transmit it [1]. Consistent with the community-based evaluation of a gamified XR prototype among the Kenyah and Kayan [1], any preservation effort must be co-designed with knowledge holders and align with free, prior, and informed consent principles. Ethical concerns also include data sovereignty, ownership of digitized plant and practice records, and the long-term stewardship of XR assets after a project concludes [6]. From an epistemological standpoint, XR tends to privilege visual and spatial modalities; this may be a good fit for plant identification and gesture-based preparation, but it can flatten the narrative, sensory, and relational dimensions of IMK. A notable proportion of the 23 reviewed studies derive from general XR applications rather than from IMK-specific contexts; this review therefore synthesizes evidence across domains and makes its IMK inferences with appropriate caution. Public-health XR literature [2] and the broader digital-age XR review [6] converge on the view that cultural fit, equity, and ethical governance must be designed-in rather than retrofitted.

V. CONCLUSION AND FUTURE WORKS

This systematic literature review brought together findings from 23 peer-reviewed studies published from 2019 to 2026 on the use of Extended Reality in educational and cultural settings [1], [2], [5], [6]. The analysis shows that XR technologies, including AR, VR, and MR, have gained considerable use in

improving learning outcomes, visualizing cultural heritage, and encouraging public engagement [1], [2], [5], [6], [8], [9]. The key finding of this review is that gamification works well as a teaching tool within XR environments, where it consistently improves intrinsic motivation and knowledge retention [1], [5], [21].

The review identifies five key points for XR implementation. First, XR shows strong potential for improving engagement through immersive learning environments [2], [5], [17]. Second, gamification supports community-based learning by creating participatory environments that improve educational value [1], [5], [21]. Third, long-term success depends on involving diverse community members throughout the process [1]. Fourth, mobile and low-cost XR platforms are the most practical choice for community settings [1], [6]. Fifth, all digitalization efforts should be guided by ethical standards that support equity and inclusivity [6].

For future academic work, studies should move beyond general educational contexts and focus specifically on knowledge preservation in community settings [1], [6]. Empirical and field-based research is needed to assess how well XR and gamified learning approaches work in real community environments [1], [6]. Development efforts should also prioritize mobile-compatible XR systems to improve accessibility in low-resource settings. Ethical frameworks should be embedded into project design to ensure transparency and long-term cultural protection [1], [6].

Looking forward, the research agenda points to several concrete directions. First, community-anchored field trials of mobile-compatible XR prototypes should be conducted with Indigenous knowledge holders, with outcomes measured beyond usability to include cultural continuity and intergenerational transfer indicators [1]. Second, the Task-Technology Fit framework applied in this review should be extended into an evaluation instrument that can be quantitatively tested in future IMK-XR studies [4], [24]. Third, hybrid AR-plus-gamification designs warrant dedicated investigation for low-bandwidth rural deployment [1], [5], [21]. Fourth, longitudinal studies are needed to assess whether XR-mediated preservation sustains knowledge retention and community engagement over years rather than single sessions. Finally, future work should integrate ethical and data-sovereignty governance into the XR development lifecycle, supported by the public-health and digital-age XR literature [2], [6], so that preservation efforts remain accountable to the communities whose knowledge they document.

ACKNOWLEDGMENT

This research was supported by the Universiti Malaysia Sarawak (UNIMAS), under the internal research grant IRG/F08/SDEC/86654/2025. The funding body was not involved in the design, data collection, analysis, or interpretation of the study.

DECLARATION OF AI ASSISTANCE

The authors acknowledge the use of a Generative AI tool (ChatGPT by OpenAI) to support language refinement and structural editing in the preparation of this manuscript. All scientific ideas, arguments, results, and conclusions were developed solely by the authors. The authors take full responsibility for the accuracy, originality, and integrity of the content.

REFERENCES

[1] A. D. Hendrick et al., "Gamified extended reality potential for Kenyah and Kayan traditional medicine knowledge transfer and preservation," in 2025 7th Asia Symposium on Image Processing (ASIP), IEEE, 2025, pp. 153–158, doi: 10.1109/ASIP65980.2025.00035.

[2] Y. C. J. Lee and B. P. Takenaka, "Extended reality as a means to enhance public health education," Frontiers in Public Health, vol. 10, p. 1040018, 2022, doi: 10.3389/fpubh.2022.1040018.

[3] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, 2021, doi: 10.1136/bmj.n71.

[4] D. L. Goodhue and R. L. Thompson, "Task-technology fit and individual performance," MIS Quarterly, vol. 19, no. 2, pp. 213–236, 1995, doi: 10.2307/249689.

[5] J. Cross et al., "Using extended reality in flight simulators: a literature review," IEEE Trans. Vis. Comput. Graph., vol. 29, no. 9, pp. 3961–3975, 2023, doi: 10.1109/TVCG.2022.3173921.

[6] E. Ottaviano, "Extended reality in the digital age: a literature review," Computers & Education: X Reality, vol. 8, p. 100127, 2026, doi: 10.1016/j.cexr.2025.100127.

[7] R. Chen and B. Liao, "Application of XR technology in stomatology education: theoretical basis, application scenarios and future prospects," in Proc. 2023 IEEE Conf. Virtual Reality and 3D User Interfaces Workshops and Short Papers (VRW), Shanghai, China, 2023, pp. 108–111, doi: 10.1109/VRW58643.2023.00108.

[8] S. Barteit et al., "Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review," JMIR Serious Games, vol. 9, no. 1, p. e13748, 2021, doi: 10.2196/13748.

[9] A. Pregowska et al., "Information and communication technologies combined with mixed reality as supporting tools in medical education," Electronics, vol. 11, no. 3, p. 369, 2022, doi: 10.3390/electronics11030369.

[10] Y. Li et al., "The next generation of healthcare ecosystem in the metaverse," Biomedical Journal, vol. 47, no. 2, pp. 149–158, 2024, doi: 10.1016/j.bj.2023.12.001.

[11] T. Nanayakkara et al., "Evaluating feature-based homography pipelines for dual-camera registration in acupoint annotation," Journal of Imaging, vol. 11, no. 1, p. 8, 2025, doi: 10.3390/jimaging11010008.

[12] J. S. Cho, R. Jotwani, S. Chan, D. M. Thaker, J. D. On, R. J. Yong, and D. Hao, "Extended reality navigation for pain procedures: a narrative review," Reg. Anesth. Pain Med., 2024, doi: 10.1136/rapm-2024-105352.

[13] P. Pons et al., "Extended reality for mental health: current trends and future challenges," Frontiers in Computer Science, vol. 4, p. 1034307, 2022, doi: 10.3389/fcomp.2022.1034307.

[14] T. Kim et al., "Inclusive user experience design for older adults focusing on XR-based cognitive and physical training," in Proc. 2024 IEEE Int. Conf. Big Data and Smart Computing (BigComp), 2024, pp. 436–438, doi: 10.1109/BigComp60711.2024.00096.

[15] B. Wells, N. Wang, S. Wiebe, C. B. Josephson, F. Sinaei, and F. Maurer, "NeuroCR: integrating mixed reality with desktop medical imaging software for pre-surgical evaluations of epilepsy patients," in Proc. 2025 IEEE Conf. Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 1236–1237, doi: 10.1109/VRW66409.2025.00264.

[16] G. Koutalieris et al., "Enhancing urban environmental sustainability through unified stakeholders needs co-creation process (AENEA)," in Proc. 2023 IEEE Int. Conf. Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), Milano, Italy, Oct. 2023, doi: 10.1109/METROXRAINE58569.2023.10405595.

[17] S. K. T. Bailey et al., "Immersive distance simulation: exploring the educational impact of stereoscopic extended reality (XR) video in remote learning environments," Medical Teacher, vol. 46, no. 9, pp. 1134–1136, 2024, doi: 10.1080/0142159X.2024.2314725.

[18] C. Stramba-Badiale et al., "Spatial memory rehabilitation using low-cost virtual reality in patients with mild cognitive impairment: methodological proposal," in Proc. 2025 IEEE Int. Conf. Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), 2025, pp. 949–954, doi: 10.1109/METROXRAINE66377.2025.11339997.

[19] S. K. Donovan et al., "Expansion of simulation and extended reality for interprofessional education," J. Interprofessional Educ. Pract., vol. 22, p. 100421, 2021, doi: 10.1016/j.xjep.2021.100421.

[20] M. del Mar Uclés Torrente et al., "Exploring digital twin applications in pediatric anthropometry: a preliminary validation study," in Proc. 2025 IEEE Int. Conf. Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), pp. 1337–1342, doi: 10.1109/METROXRAINE66377.2025.11340053.

[21] F. M. Fung et al., "Digital futures of learning pharmacology and medicinal and organic chemistry," ACS Pharmacology & Translational Science, vol. 7, no. 3, pp. 1191–1194, 2024, doi: 10.1021/acsptsci.4c00043.

[22] E. A. Merlinsky, A. Gavric, H. Stojkovic, and E. Zivanovic, "Physics-driven methods for adaptive optics effect in extended reality: eliminating the need for corrective lenses," in Proc. 2023 IEEE 33rd Int. Conf. Microelectronics (MIEL), 2023, doi: 10.1109/MIEL58498.2023.10315935.

[23] V. Mancuso, S. N. Saritas, E. Pedroli, F. Franchini, S. G. Di Santo, S. Arlati, and M. Cavallo, "From laboratory to real life: psychometric properties of a VR-based ecological memory assessment tool," in Proc. 2025 IEEE Int. Conf. Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), pp. 943–948, doi: 10.1109/METROXRAINE66377.2025.11340260.

[24] D. Ratmono et al., "Virtual reality and perceived learning effectiveness in accounting studies: the mediating role of task-technology fit," Cogent Business & Management, vol. 11, no. 1, p. 2316890, 2024, doi: 10.1080/23311975.2024.2316890.

IFDA-EMA-YOLOv9: Wheat Disease Detection Integrating Flow Optimization and Auxiliary Supervision

Jiaxiang Fan, Leixiao Li*
College of Intelligent Science and Technology, Inner Mongolia University of Technology, Hohhot, China

***Abstract*—To address the challenges of feature extraction in complex field environments, the limited sensitivity of YOLOv9 to subtle disease features, and the lack of adaptive hyperparameter optimization, this paper proposes an improved high-precision detection model, named IFDA-EMA-YOLOv9. First, to enhance feature extraction capabilities, an Efficient Multi-scale Attention (EMA) mechanism and residual connections are incorporated into the network architecture. This integration effectively suppresses background noise interference and significantly improves the model's ability to aggregate and represent multi-level features of wheat disease lesions. Second, to tackle localization deviations caused by the irregular geometric shapes of disease lesions, an auxiliary box mechanism is integrated into the Complete Intersection over Union (CIoU) loss function, optimizing the regression process to improve the fit of detection boxes. Furthermore, an Improved Flow Direction Algorithm (IFDA) is employed to perform global optimization of the critical model hyperparameters, thereby avoiding the blindness of manual tuning and the local optimum trap. Experimental results on the LWDCD2020 dataset demonstrate that the proposed IFDA-EMA-YOLOv9 significantly outperforms current state-of-the-art (SOTA) methods, achieving substantial improvements in Precision, Recall, and mAP@0.5 by 6.4%, 5.94%, and 6.66%, respectively. These results demonstrate the effectiveness and robustness of the proposed method for wheat disease and pest detection.**

Keywords*—Wheat disease detection; Object detection; YOLOv9; Flow algorithm

I. Introduction

Wheat is one of the world's major staple crops and plays a crucial role in ensuring food security and sustainable agricultural production. However, throughout the full growth cycle of wheat, the frequent occurrence of major pests and diseases such as yellow rust, fusarium head blight, and aphids has become a key factor restricting yield and quality. Statistics indicate that in years with prevalent diseases, the aforementioned disasters can cause wheat yield reductions of 20% to 50%, and in extreme cases, even trigger total seasonal harvest failure, bringing inestimable losses to agricultural production. Therefore, constructing a precise, real-time automated monitoring system possesses significant practical significance for assisting farmers in making scientific decisions, implementing precise pesticide application, and improving pesticide utilization rates.

For a long time, traditional disease monitoring has mainly relied on manual field inspections or simple physical instrument detection. These methods are not only time-consuming and labor-intensive with high subjectivity, but also difficult to adapt to the monitoring demands of large-scale intensive planting, suffering from serious lag. With the rapid development of artificial intelligence technology, utilizing computer vision and image processing technology to realize rapid, non-destructive identification of crop diseases has become an innovative direction in the field of smart agriculture. Compared with traditional means, deep learning-based methods can automatically and efficiently extract disease features from complex backgrounds, significantly enhancing the intelligence level of prevention and control work.

Among numerous deep learning models, the YOLO (You Only Look Once) series algorithms stand out due to their efficient end-to-end detection mechanisms. In particular, the latest YOLOv9 model, by introducing Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN), has achieved excellent performance in core computer vision tasks such as general object detection, image classification, and semantic segmentation, demonstrating superior feature extraction capabilities and real-time inference performance [1-3].

Although YOLOv9 performs strongly on general datasets, it still faces numerous challenges that urgently need to be resolved when directly deployed to complex, unstructured wheat field environments: First, the field environment has variable illumination and complex backgrounds, so the applicability and robustness of YOLOv9 in specific agricultural scenarios require in-depth investigation and validation; second, the model's original pooling dimensionality reduction strategy, when modeling cross-channel long-range dependency relationships, easily causes the loss of subtle disease features, thereby damaging the integrity of visual representation; third, the existing CIoU loss function finds it difficult to achieve precise bounding box regression when processing targets like wheat disease lesions, which have extremely variable aspect ratios and irregular geometric shapes; finally, the configuration of model hyperparameters often relies on manual experience and lacks adaptive optimization schemes for specific datasets, limiting the upper bound of model performance.

II. Related Work

A. Progress in Wheat Pest and Disease Detection Technology

In recent years, although the YOLO series has made significant progress in the agricultural field, the research focus has shifted from simple architecture stacking to refined feature

*Corresponding author.

modeling. Qiao et al. [4] enhanced the global feature extraction capability of YOLOv8n by incorporating the C2f-iRMBS module and Inner-EIoU loss function, reaching a mAP of 96.3%. However, addressing the problem of feature redundancy in field environments, the latest research trends have begun to introduce Partial Convolution (PConv) to extract more representative spatial features by reducing redundant calculations. Furthermore, considering the irregularity of wheat lesion morphology, the Large Selective Kernel mechanism has demonstrated advantages in capturing multi-scale lesion features by dynamically adjusting the receptive field. In this context, this paper proposes the EMA-YOLOv9 model, which, by integrating the Efficient Multi-Scale Attention (EMA) mechanism with a residual architecture, aims to solve the negative impact of cross-channel relationship modeling on visual representation under complex backgrounds, further reinforcing the flexibility of feature extraction.

Addressing the difficulty of locating tiny lesions, the traditional IoU loss function has limitations when processing wheat disease lesions of variable sizes. Latest technologies such as Inner-IoU [5] have accelerated convergence by introducing auxiliary bounding box scaling factors, while MPDIoU [6] has effectively solved the problem of inaccurate bounding box overlap measurement using minimum point distance. Drawing on this frontier idea, we introduce an auxiliary frame mechanism into the YOLOv9 framework to optimize the CIoU loss function, effectively balancing localization accuracy among targets of different scales.

B. YOLOv9 Parameter Optimization

YOLOv9 utilizes a reversible information callback mechanism to significantly alleviate feature loss in deep networks. However, its performance is highly dependent on complex hyperparameter configurations. Existing research on YOLOv9 hyperparameter automation remains relatively scarce, with most literature still remaining at optimization schemes for earlier versions.

Reviewing past research, YOLOv5 employed the Genetic Algorithm (GA) [7] to optimize hyperparameters, and Mahmudah et al. [8] utilized the Tree-structured Parzen Estimator (TPE) in YOLOv7 to explore the search space. However, facing the more non-linear parameter space of YOLOv9, traditional Genetic Algorithms, Artificial Bee Colony (ABC) algorithms [9], or Opposition-based Optimization Strategies (OOS) [10] are prone to falling into local optima and are difficult to cover the full hyperparameter set. To excavate the model's potential, this paper innovatively adopts the Improved Flow Direction Algorithm (IFDA) for global optimization and proposes utilizing Chaotic Mapping and Lévy Flight mechanisms to enhance population diversity; the final IFDA-EMA-YOLOv9 model realizes the leap of hyperparameter configuration from empirical tuning to automated evolution.

III. Basic Concepts

A. YOLOv9

YOLOv9 employs a regression strategy to achieve end-to-end simultaneous prediction for target region identification, spatial localization, and category classification, striking a significant balance between detection accuracy and inference efficiency. The overall architecture of the algorithm comprises four core components: the Backbone, the Auxiliary Reversible Branch, the Multi-level Auxiliary Information, and the Head. During the model training phase, YOLOv9 generates a total of six sets of output feature maps for subsequent prediction label inference. Distinct from previous architectures, YOLOv9 innovatively introduces an auxiliary output module. By calculating the correlation between the prediction loss and ground truth labels, this mechanism facilitates the efficient back-propagation of gradient information through the auxiliary reversible path to shallow feature extraction modules. This process optimizes the gradient backflow from deep networks to shallow networks, which is crucial for the effective updating of backbone weights and the maintenance of feature representation capabilities. The core task of the model lies in accurately predicting the spatial locations of targets of different sizes based on multi-scale feature maps. By integrating the Silence module to optimize the inference process, combining standard CBS (Conv-BN-SiLU) modules, and refining the loss function and optimizer strategy, the model achieves significantly enhanced robustness in object detection and localization. Comparative experiments demonstrate that, relative to previous versions such as YOLOv7 and YOLOv8, YOLOv9 achieves improved detection accuracy while significantly reducing the number of parameters and computational complexity (FLOPs). In summary, YOLOv9 represents significant technical progress in the trade-off between processing speed and detection performance.

B. Flow Direction Algorithm (FDA)

1) Algorithm background and d8 mechanism: The Flow Direction Algorithm is inspired by natural surface runoff processes. In hydrology, direct runoff is driven by gravity and converges toward the watershed outlet along the direction of the terrain slope. The classical method for simulating this process involves discretizing the watershed into regular grid cells and determining the runoff path by calculating the elevation difference and slope between rasters. Among these methods, the D8 algorithm (Deterministic-8 Node) proposed by O'Callaghan et al [11]. is currently the mainstream method for determining flow direction. In the D8 model, it is assumed that each central raster has eight adjacent rasters. By calculating the slope between the central raster and its neighborhood rasters, the flow direction is defined as pointing to the adjacent raster with the maximum slope descent (as shown in Fig. 1).

2) Assumptions of FDA: Drawing on the physical mechanism of the D8 algorithm, FDA abstracts the solving process of optimization problems into the process of rainfall converting into runoff and converging toward the outlet. In this mapping system, water flow represents population individuals, position represents the solution vector, and elevation corresponds to the objective function value. Fig. 2 illustrates the flowchart of FDA. Based on literature [12], the operation of FDA is based on the following core assumptions:

- Individual definition: Each flow possesses determined position coordinates and a corresponding elevation (fitness value) in the solution space.

- Neighborhood structure: Each flow possesses β adjacent positions in the space, and each neighborhood position corresponds to an elevation or objective function value.
- Dynamic relationship: The flow velocity of the fluid is proportional to the path slope.
- Velocity definition: The velocity of the fluid moving toward the lowest elevation (the direction of the optimal value) is defined as V.
- Convergence goal: The position of the watershed outlet corresponds to the global optimal solution of the objective function in the search space.

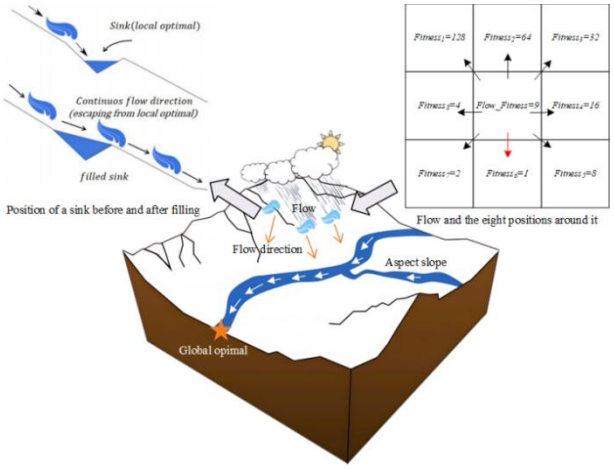


Fig. 1. D8 Method process flow diagram.

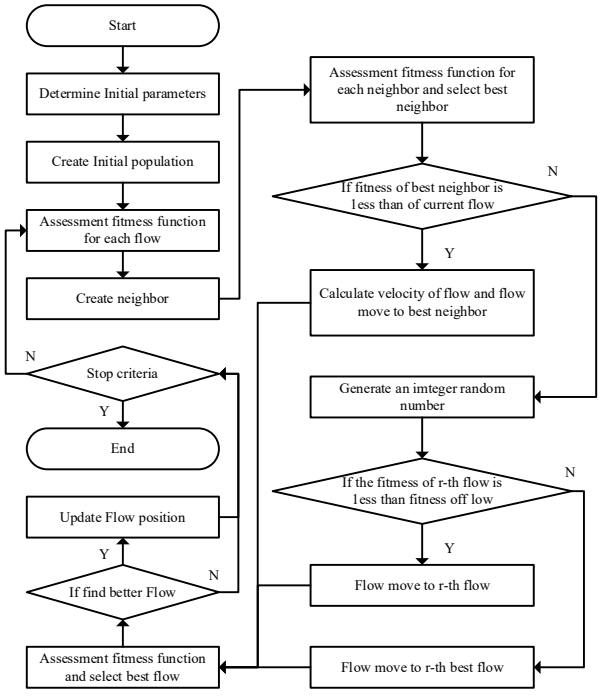


Fig. 2. Flowchart of the original FDA.

3) Initialization of flow positions: The initialization phase of the FDA algorithm involves the setting of key parameters such as population size ($Size_{pop}$), neighborhood quantity ($Size_{near}$), neighborhood radius (R_{near}), and maximum iterations (MaxIter). The initial positions of various fluids in the population are randomly generated within the solution space via the following equation:

$$Flow_X(i) = lb + r_1 \times (ub - lb) \tag{1}$$

Where $Flow_X(i)$ denotes the position of the i-th flow, lb and ub represent the lower and upper bounds of the decision variables, respectively; and r_1 is a random number uniformly distributed in [0,1].

4) Neighborhood generation strategy: FDA assumes that a neighborhood with a radius of R_{near} exists around each flow. Its position is generated by the following formula:

$$Neighbor_X(i) = Flow_X(i) + r_2 \times R_{near} \tag{2}$$

Here, denotes the position of the j-th neighbor. represents a random number drawn from a normal distribution with a mean of 0 and a standard deviation of 1. A smaller value of restricts the search to a narrow range, whereas a larger value allows for exploration across a broader scope. Searching within a wide range yields more diverse solutions and enhances the probability of identifying near-optimal solutions through global search.

Therefore, FDA adopts a formula to make the value of R_{near} linearly decrease from large to small, so as to achieve a balance between global search and local search. The mathematical formula is as follows:

$$\begin{aligned} R_{near} &= (r_3 \times Rand_X - r_3 \times Flow_X(i)) \\ &\times \left\| Best_X - Flow_X(i) \right\| \times W \end{aligned} \tag{3}$$

Where, r_3 is a uniformly distributed random number; $Rand_X$ is a random position generated by (1); W is a non-linear weight that varies randomly between 0 and infinity. The first term of (3) indicates that $Flow_X(i)$ moves towards the random position ($Rand_X$). As the number of iterations increases, $Flow_X(i)$ converges towards $Best_X$, causing the Euclidean distance between $Best_X$ and $Flow_X(i)$ to decrease to zero. Consequently, the local search is optimized. The calculation formula for W is as follows:

$$W = ((1 - \frac{iter}{MaxIter})^{2 \times r_4}) \times (\bar{r_1} \times \frac{iter}{MaxIter}) \times \bar{r_1} \tag{4}$$

Where *iter* represents the current iteration number; r_4 is a uniformly distributed random number; and r is a uniformly distributed random vector.

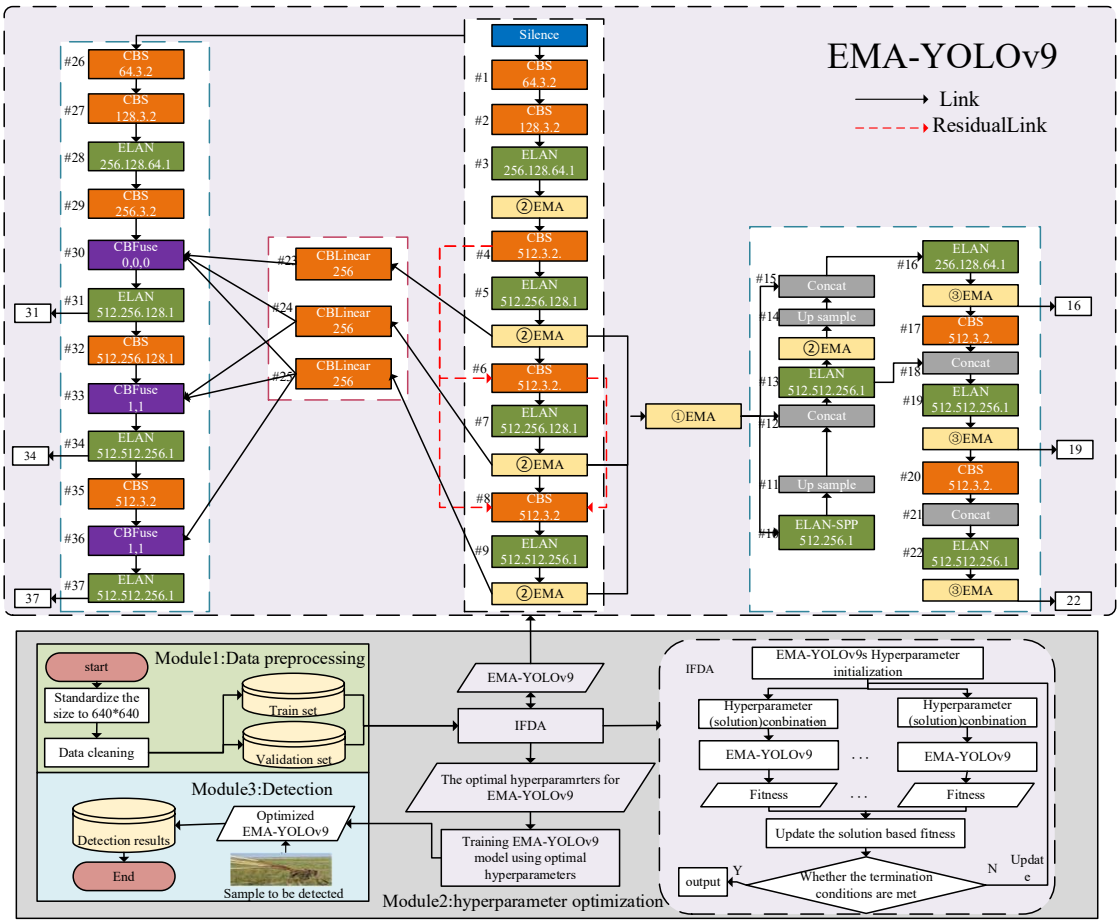


Fig. 3. Overall framework of methods for detecting wheat diseases.

5) Updating flow position: The flow moves at velocity V towards the neighbor with the optimal objective value. V is directly proportional to its slope. The calculation of V is as follows:

$$V = r_5 \times S \tag{5}$$

Where r_5 is a uniformly distributed random number; S represents the slope vector from the neighbor to the current flow position. S for the i-th flow relative to the j-th neighbor is also determined by the following formula:

$$S(i,j,d) = \frac{fitness(Flow_X(i,d)) - fitness(Neighbor_X(j,d)))}{\left\| Flow_X(j,d) - Neighbor_X(j,d) \right\|} \tag{6}$$

Where fitness() represents the function for calculating the objective value; and d represents the problem dimension.

The new position of the flow $Flow_X(i)$ is calculated by the following formula:

$$Flow_newX(i) = Flow_X(i) + V \times \frac{Flow_X(i) - Neighbor_X(j)}{\left\| Flow_X(i) - Neighbor_X(j) \right\|} \tag{7}$$

The function value must not be lower than the objective function value of the flow. This concept is analogous to the water-filling process of a sink, which determines the direction of the flow. To simulate the aforementioned state, the FDA randomly selects another flow using the following equation:

$$Flow_newX(i) = \begin{cases} Flow_X(i) + \overline{r_2} \times (Flow_X(r) - Flow_X(i)) \\ fitness(Flow_X(r)) < fitness(Flow_X(i)) \\ Flow_X(i) + r_6 \times (Best_X - Flow_X(i)) \\ otherwise \end{cases} \tag{8}$$

Where, r is a random integer; r_2 is a random vector uniformly distributed in $[0, 1]$; and r_6 is a uniformly distributed random number. Finally, the fitness value of the newly generated flow is calculated and compared with that of the previous flow to determine its position at iteration Iter+1.

IV. Methodology

A. General Framework

As illustrated in Fig. 3, the comprehensive methodology for identifying wheat diseases and pests utilizing the IFDA-EMA-YOLOv9 approach comprises two primary components:

network architecture design and full-process optimization strategies.

At the network architecture level, the model is built upon YOLOv9 as the baseline. The backbone and neck networks are constructed by integrating CBS, ELAN, and CBFuse modules. Notably, the Enhanced Multi-scale Attention (EMA) mechanism module is introduced to bolster the model's capability in capturing key features and to improve the efficiency of multi-scale feature fusion.

At the system workflow level, the framework establishes a closed-loop processing mechanism: first, Module 1 performs preprocessing tasks, including image dimension standardization and data cleaning; subsequently, Module 2 integrates IFDA into EMA-YOLOv9 for iterative hyperparameter optimization to achieve the optimal model configuration; finally, Module 3 utilizes the optimized model to execute high-precision object detection tasks on the input samples.

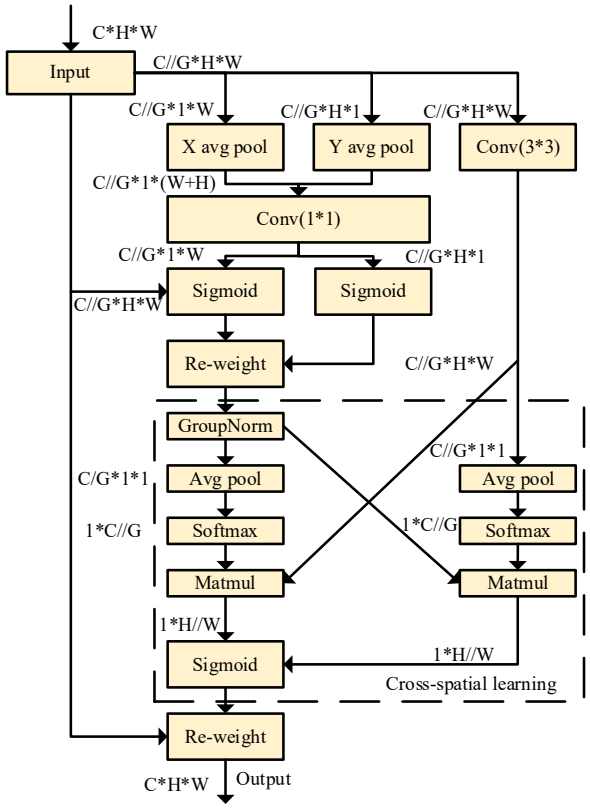


Fig. 4. Structure of EMA.

B. EMA-YOLOv9

1) Multi-level EMA: The relatively small size of wheat and its leaves, coupled with colors that are easily confused with other foliage, poses challenges for traditional YOLO detection models. Incorporating attention mechanisms has emerged as a viable solution to address these issues. IFDA-EMA-YOLOv9 introduces an enhanced multi-scale attention mechanism into YOLOv9, which facilitates the uniform distribution of spatial semantic features across each feature map. The multi-scale architecture of EMA is particularly effective in establishing short-term and long-term dependencies, thereby strengthening the model's feature extraction capabilities and ultimately enhancing object detection performance [13]. The structural representation of EMA is shown in Fig. 4.

EMA is defined as an efficient multi-scale attention mechanism characterized by the absence of dimensionality reduction and its role as a cross-space learning method. It establishes short-term and long-term dependencies through a multi-scale parallel sub-network. By modeling long-term dependencies, the EMA mechanism preserves precise positional information. By fusing contextual information across different scales, the object detection model gains the ability to generate pixel-level attention for feature maps. Compared to traditional Coordinate Attention, EMA utilizes only two convolution kernels (1×1 and 3×3). To aggregate multi-scale spatial structural information, the 3×3 kernel is arranged in parallel with the 1×1 branch to facilitate rapid response. This parallel sub-structure allows the network to avoid extensive sequential and deep processing. EMA enables the learning of effective channel representations without the need for channel dimensionality reduction, thereby generating superior pixel-level attention for high-level feature maps.

As shown in Fig. 5, three methods have been designed to integrate EMA into YOLOv9, including implementations at the backbone and neck/head levels:

Placing EMA at the end of the backbone, as shown in Fig. 5 (a).

Placing EMA after the RepNCSSPELAN4 module in the backbone, as shown in Fig. 5 (b).

Embedding EMA after the RepNCSSPELAN4 module within the Neck and Head structures, as shown in Fig. 5 (c).

This paper aims to investigate the performance of multi-level EMA through experiments and, based on the three aforementioned implementation strategies, determine the optimal configuration for the integration of YOLOv9 and EMA.

2) ICIoU Loss: The original YOLOv9 model employs CIoU (Complete Intersection over Union) loss as its loss function. CIoU loss comprehensively considers three geometric parameters: overlap area, central point distance, and aspect ratio [14]. The calculation formula for CIoU loss is as follows:

$$CIoU = IoU - \left(\frac{\rho^2(b, b^{gt})}{c^2} + \kappa\vartheta\right) \tag{9}$$

Where IoU represents the overlapping area; $\rho^2(b, b^{gt})$ represents the square of the Euclidean distance between the center points of the predicted box and the ground truth box; c^2 represents the square of the diagonal length of the smallest enclosing rectangle covering both boxes; κ and ϑ are aspect ratio parameters, calculated as follows:

$$\kappa = \frac{\vartheta}{1 - IoU + \vartheta} \tag{10}$$

$$\vartheta = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}) \tag{11}$$

Here, w^{gt} and h^{gt} represent the height and width of the ground truth box, while h and w represent the height and width of the predicted box. The CIoU loss function accounts for both central point distance and the consistency of the aspect ratio. However, the variability in the shapes and scales of wheat and its leaves limits the effectiveness of CIoU loss in accurately calculating IoU for irregularly shaped targets. Furthermore, CIoU loss struggles to maintain stable localization accuracy across different target scales, thereby impacting the overall detection accuracy

To address the aforementioned challenges, we combine the Inner IoU with the CIoU loss and introduce an Auxiliary Box into the CIoU framework, thereby proposing the ICIoU loss function. The structural diagram of the auxiliary box is illustrated in Fig. 6.

The center point of the auxiliary box coincides with that of the ground truth box, but their dimensions differ. The auxiliary box serves as a representation of the ground truth box. In cases of high IoU values, the absolute value of the IoU gradient for a smaller-scaled auxiliary box is greater than that of the ground truth box. Conversely, for samples with low IoU values, the absolute value of the IoU gradient for a larger-scaled auxiliary box is smaller than that of the ground truth box. Therefore, adjusting the dimensions of the auxiliary box contributes to accelerating model convergence.

By adjusting the parameter r_{IoU}, the dimensions of the auxiliary box can be customized to adapt to different datasets. This design facilitates more precise fine-tuning of the predicted box position when analyzing wheat pest images with high IoU values. Conversely, when processing images with low IoU values, a larger auxiliary box can be employed to fully cover the target region, thereby mitigating localization errors associated with CIoU loss.

The calculation method for the auxiliary box is described as follows. First, the four boundary coordinates of the actual auxiliary box must be calculated respectively:

$$b_t^{gt} = y_c^{gt} - \frac{h^{gt} \times r_{IoU}}{2} \tag{12}$$

$$b_b^{gt} = y_c^{gt} + \frac{h^{gt} \times r_{IoU}}{2} \tag{13}$$

$$b_l^{gt} = x_c^{gt} - \frac{w^{gt} \times r_{IoU}}{2} \tag{14}$$

$$b_r^{gt} = x_c^{gt} + \frac{w^{gt} \times r_{IoU}}{2} \tag{15}$$

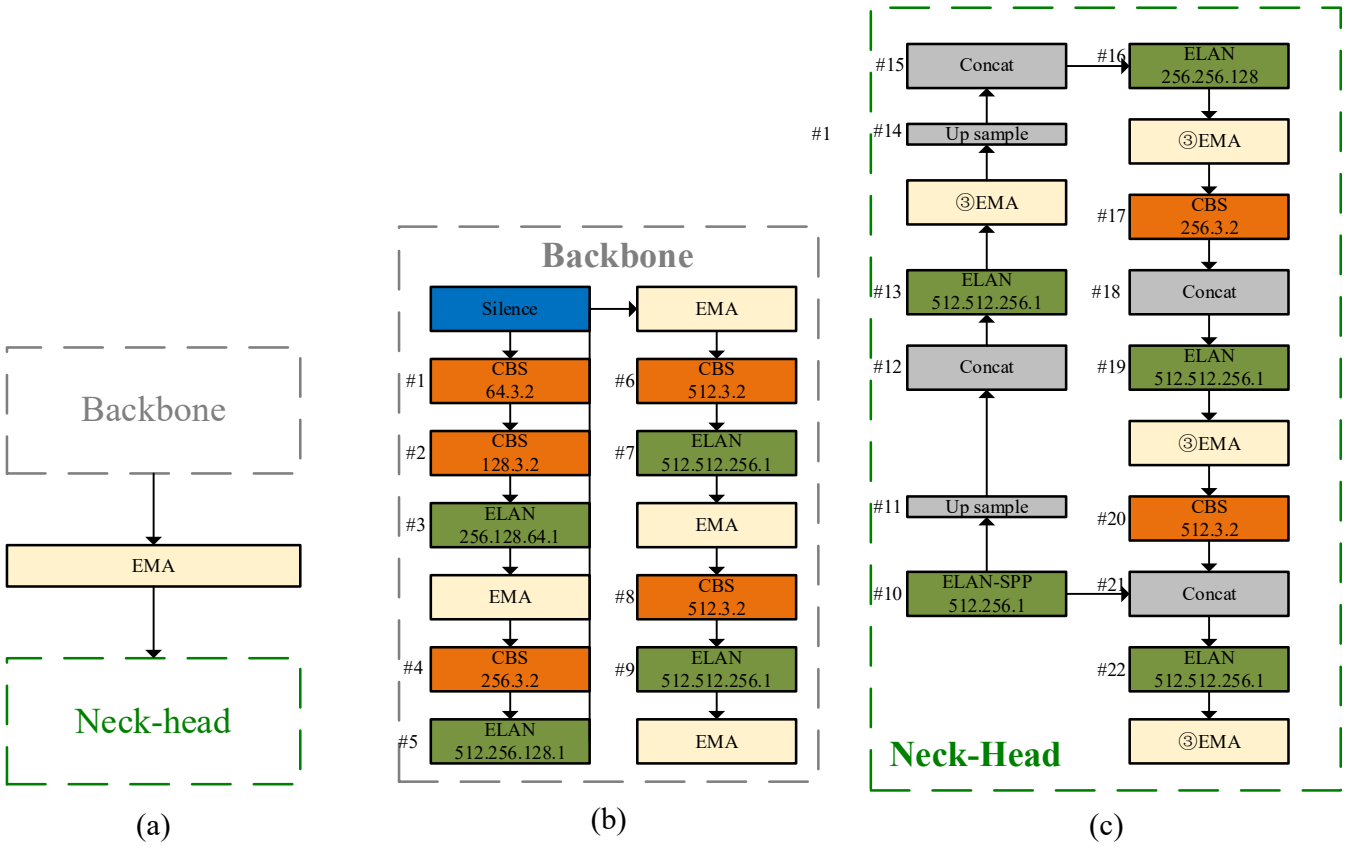


Fig. 5. Diagrammatic sketch of multi-level EMA.

Similarly, the four boundary coordinates of the predicted auxiliary box— b_t 、 b_b 、 b_l and b_r —are calculated respectively. Subsequently, the intersection and union of the auxiliary boxes are calculated:

$$inter = (min(b_r^{gt}, b_r) - max(b_l^{gt}, b_l)) \times (min(b_b^{gt}, b_b) - max(b_t^{gt}, b_t)) \quad (16)$$

$$union = (w^{gt} h^{gt})(r_{IoU})^2 + (wh)(r_{IoU})^2 - inter \quad (17)$$

Subsequently, the intersection-over-union (IoU) term $1 - \frac{inter}{union}$ of the auxiliary box is calculated to serve as the auxiliary box loss. Finally, the loss function of EMA-YOLOv9 is expressed as:

$$ICIoU = IoU - (\frac{\rho^2(b, b^{gt})}{c^2} + \kappa\upsilon) + 1 - \frac{inter}{union} \quad (18)$$

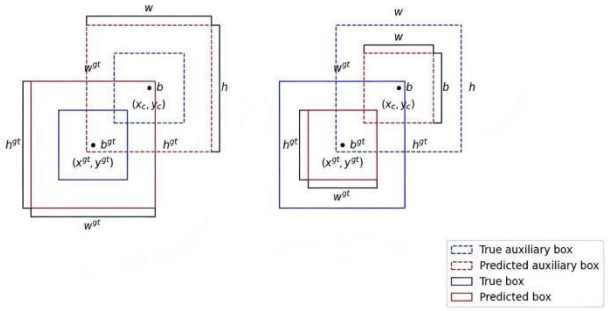


Fig. 6. Schematic diagram of auxiliary frame.

3) Cross-scale feature interaction: To enhance the feature extraction capability of YOLOv9, residual connections were introduced into the backbone network architecture of this model. In the field of object detection, it is generally recognized that low-level features contain richer edge and texture details, whereas high-level features tend to represent more abstract semantic content. Cross-level feature fusion is considered beneficial for improving the model's performance in tasks such as small object detection and fine segmentation. The implementation of residual connections enables YOLOv9 to effectively capture deep information that is difficult to adequately express with single-level features, thereby enriching the hierarchical representation of features and facilitating the acquisition of high-level semantic information. IFDA-EMA-YOLOv9 adopts a residual structure derived from the ResNet architecture, establishing connections among the 4th, 6th, and 8th layers of the backbone network to promote cross-scale feature interactions. The specific residual structure employed is illustrated in Fig. 7. This structure initially utilizes a 1x1 convolution to perform dimensionality reduction on the residual connection, subsequently connects to a 3x3 convolution, and finally increases the dimensionality through another 1x1 convolution.

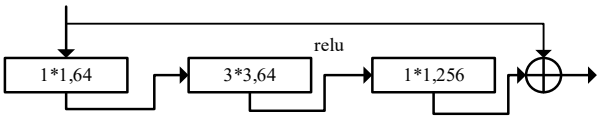


Fig. 7. Matic diagram of residual structure.

C. IFDA

The Improved Flow Direction Algorithm (IFDA) introduces three enhancements based on the original FDA, specifically encompassing solution initialization, the fitness function, and the solution update mechanism.

Meanwhile, to achieve high-precision training of deep learning models, hyperparameters must be adjusted. Notably, evaluating each parameter set requires the model to complete a full training cycle. The formal description of the parameter tuning problem is as follows:

$$\min errorRate = F(P) \quad s.t. \begin{cases} P = \{p_1, p_2, \ldots, p_i\} \\ p_i \in [\mu b_i, lb_i] \\ i = \{1, 2, \ldots, k\} \end{cases} \quad (19)$$

Where $errorRate$ denotes the error of the trained model; P represents the set of model hyperparameters; k indicates the number of hyperparameters requiring parameterization; and $[ub_i, lb_i]$ signifies the variation range of parameter p_i.

Early methodologies primarily relied on manual tuning, an approach heavily dependent on the practitioners' expert experience. However, given the inherent uncertainty and time-consuming nature of manual tuning, grid search techniques were gradually adopted. Nevertheless, in the training of deep learning models characterized by numerous parameters and extensive value ranges, grid search often results in high computational complexity and excessive time consumption. This complexity stems from the significantly higher training overhead of deep learning models compared to shallow models such as Support Vector Machines (SVM). Consequently, this paper proposes utilizing heuristic algorithms for parameter optimization.

1) Solution initialization: IFDA must first randomly generate initial solutions within the constrained solution space. This paper summarizes all hyperparameters that may significantly impact the performance of EMA-YOLOv9, as shown in Table I.

The hyperparameters requiring further explanation are as follows:

- Determines the model version of YOLOv9. The original authors of YOLOv9 proposed five different versions, namely YOLOv9-T, YOLOv9-S, YOLOv9-M, YOLOv9-C, and YOLOv9-E, each corresponding to a different number of parameters. Specifically, the YOLOv9-E version contains 57.3 million parameters, while the YOLOv9-T version has the fewest, at only 2 million.

TABLE I. HYPER-PARAMETERS OF EMA-YOLOV9 THAT NEED TO BE OPTIMIZED

Category	Parameter	Symbol	Value range
Model Structure Hyperparameters	Model size	$Size$	$Size \in \{0,1,2,3,4\}$
	Location of EMA	α_{EMA}	$\alpha_{EMA} \in \{0,1,2,3,4,5\}$
	Network width gain	d_m	$d_m \in (0,1]$
	Network width gain	w_m	$w_m \in (0,1]$
Model Training Hyperparameters	Activation function	AF	$AF \in \{0,1,2,3,4,5\}$
	Batch size	BS	$BS \in [8,256], BS \in E^+$
	Warm up epochs	$epochs_{warm}$	$epochs_{warm} \in [1,50], epochs_{warm} \in N^+$
	Initial learning rate	LR	$LR \in [0.0001, 0.001]$
	Final learning rate	LR'	$LR' \in [0.001, 0.1]$
	Momentum of optimizer	$Momentum$	$Momentum \in [0,1]$
	Weight decay coefficient	WD	$WD \in [0,1]$
	Decay rate of moment estimation index	EDR	$EDR \in [0,1]$
	IoU loss weight	w_1	$w_1 \in [0,1]$
	Weight loss of center point distance	w_2	$w_2 \in [0,1]$
	Weight loss of aspect ratio distance	w_3	$w_3 \in [0,1]$
	Tune parameter	r_{IoU}	$r_{IoU} \in [0.5,1.5]$

- α_{EMA} determines the position of the EMA module within the YOLOv9 $\alpha_{EMA}=0$ signifies that no EMA is incorporated; $\alpha_{EMA}=1$ indicates the integration of EMA after the 9th layer (#9); $\alpha_{EMA}=2$ indicates introducing EMA after layers #3, #5, #7, and #9; $\alpha_{EMA}=3$ indicates introducing EMA after layers #13, #16, #19, and #22; $\alpha_{EMA}=4$ indicates introducing EMA at layers #9, #13, #16, #19, and #22; $\alpha_{EMA}=5$ indicates introducing EMA after layers #3, #5, #7, #9, #13, #16, #19, and #22.
- d_m represents the network Depth Gain. This parameter influences the number of Bottleneck CSP modules and SPPCSPC modules within the ELAN architecture of the backbone network. This factor directly determines the overall network depth of YOLOv9.
- w_m represents the network Width Gain. This parameter influences the number of convolutional channels in both the backbone network and the detection head. Consequently, this factor directly determines the overall network width of YOLOv9.
- AF determines the activation function utilized in YOLOv9. The set $\{GeLU, ReLU, Mish, Swish, SiLU, LeakyReLU\}$ is encoded sequentially as $\{0,1,2,3,4,5\}$.
- w_1, w_2, w_3 are the three loss function weight coefficients for ICIoU, whereas the weight for the auxiliary box loss is determined by $1-w_1-w_2-w_3$.

$$ICIoU = w_1 IoU - w_2\left(\frac{\rho^2(b,b^{gt})}{c^2} + w_3\kappa\vartheta\right) + (1-w_1-w_2-w_3)\left(1-\frac{inter}{union}\right) \quad (20)$$

Traditional methods for population initialization typically involve random distribution across the entire space; however, this approach often leads to significant non-uniformity, thereby limiting diversity during the early stages of population evolution and reducing the convergence efficiency of the algorithm [15]. In light of this, we employ various Chaotic Maps to generate initial solutions for different parameters. Utilizing chaotic maps facilitates the generation of more uniformly distributed sequences, which in turn enhances population diversity.We employ the Singer chaotic map to generate the initial solutions for $Momentum$, WD, and EDR, utilize the Gaussian chaotic map to generate the initial solution for $epoch_{warmup}$, and use the Tent chaotic map to generate initial solutions for the other hyperparameters.

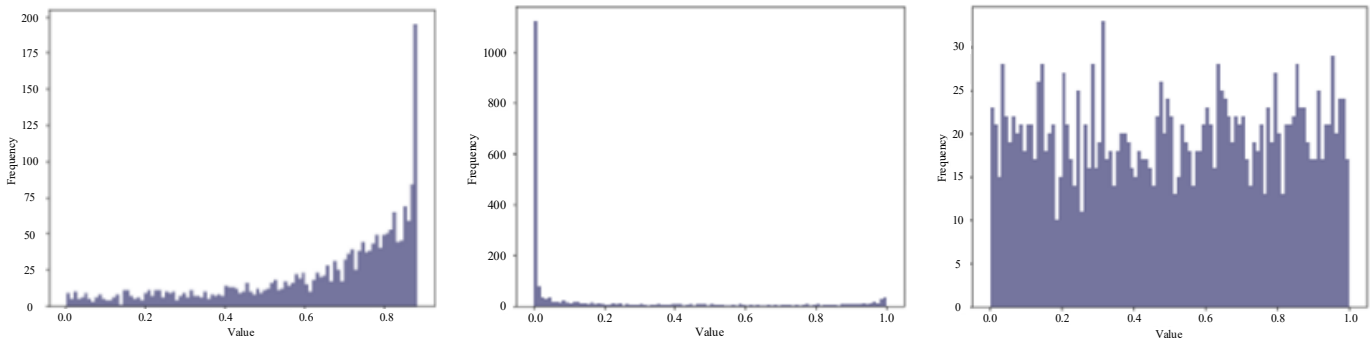


Fig. 8. Distribution curves of three chaotic maps.

The Singer chaotic map is capable of uniformly generating a relatively large number of high values, and its mapping formula is as follows:

$$x_{k+1}^{chaos} = 0.99(7.86_k^{chaos} - 23.32(x_k^{chaos})^2 + 28.75(x_k^{chaos})^3 - 13.302875(x_k^{chaos})^4) \quad (21)$$

The Gaussian chaos mapping can uniformly generate more small values; its mapping formula is:

$$x_{k+1}^{chaos} = \begin{cases} 1, x_k^{chaos} = 0 \\ \dfrac{1}{\mathrm{mod}(x_k^{chaos}, 1)}, x_k^{chaos} \neq 0 \end{cases} \quad (22)$$

The Tent Chaos Mapping can generate uniformly distributed hyperparameters; its mapping formula is:

$$x_{k+1}^{chaos} = \begin{cases} 2x_{k+1}^{chaos}, 0 \leq x_{k+1}^{chaos} \leq \dfrac{1}{2} \\ 2(1 - x_{k+1}^{chaos}), \dfrac{1}{2} \leq x_{k+1}^{chaos} \leq 1 \end{cases} \quad (23)$$

The distribution curves of the three mappings are illustrated in Fig. 8.

In this paper, mAP@0.5 is employed as the metric to evaluate the recognition performance of the model. Additionally, the number of model parameters is utilized as an indicator of model efficiency. To address the multi-objective optimization challenge of simultaneously enhancing recognition performance and reducing the number of model parameters, this paper transforms the problem into two independent single-objective optimization tasks for resolution. *Zoom(Parameters)* refers to the process of normalizing the model parameters into the [0, 1] interval based on established upper and lower bounds; these boundary values can be derived from the value ranges provided in Table I.

Furthermore, the number of additional modules in YOLOv9 also influences the total parameter count of the model. However, since the developers of YOLOv9 determined the quantity of these modules based on empirical knowledge, we did not optimize the BottleneckCSP modules, SPPCSPC modules, or other modules, except for those related to the number of convolutional channels.

2) Fitness function: The fitness function plays a crucial role in evaluating the quality of solutions and serves as the foundation of heuristic algorithms. For the hyperparameter optimization problem of EMA-YOLOv9, a fitness function as shown in (24) was constructed. This function consists of two components: recognition performance and model structural lightweighting.

$$Fitness = mAP_{0.5} - Zoom(Parameters) \quad (24)$$

We employ mAP0.5 as the metric for evaluating the recognition performance of the model. Additionally, the number of model parameters is utilized as a metric to measure model efficiency. To address the multi-objective optimization challenge of simultaneously improving recognition performance and reducing model parameters, this paper transforms it into two independent single-objective optimization problems for solution. Zoom (Parameters) refers to the process of normalizing model parameters to the [0, 1] interval based on established upper and lower bounds; these boundary values can be derived from the value ranges provided in Table I.

Furthermore, the number of additional modules in YOLOv9 also affects the total parameter count of the model. However, considering that the developers of YOLOv9 determined the quantity of these modules based on empirical knowledge, we did not optimize the BottleneckCSP modules, SPPCSPC modules, or other modules, with the exception of those related to the number of convolutional channels.

3) Update mechanism: The process of simulating convergence towards the current optimal flow may cause the algorithm to fall into local optima. Consequently, the Lévy Flight mechanism was introduced, as described in (8). Lévy Flight is characterized by a random walk that employs a heavy-tailed probability distribution for step sizes. This distribution grants a relatively high probability of occurring substantial steps during the random walk process.

Following the introduction of the Lévy Flight mechanism, the update formula for is reformulated as:

$$Flow_newX(i) = \begin{cases} Flow_X(i) + \overline{r_2} \times (Flow_X(r) - Flow_X(i)) \\ fitness(Flow_X(r)) < fitness(Flow_X(i)) \\ Flow_X(i) + r_{levy} \times (Best_X - Flow_X(i)) \\ otherwise \end{cases} \tag{25}$$

Where represents the Lévy random step size, which is calculated as follows:

$$r_{levy} = \frac{\mu_{levy}}{|v_{levy}|^{\frac{3}{2}}} \tag{26}$$

$$\mu_{levy} \sim N(0, \sigma^2). v_{levy} \sim N(0,1)$$

$$\sigma = \left[\frac{\Gamma(1+1.5)\sin(\frac{3\pi}{4})}{1.5\Gamma(\frac{1+1.5}{2})2^{\frac{0.5}{2}}} \right]^{\frac{3}{2}} \tag{27}$$



Fig. 9. Example image of the LWDCD2020 dataset.

V. Experiments

A. Experimental Details

In this experimental study, the Adam optimizer is utilized within the YOLOv9 model, with the number of training epochs set to 300. The default values for momentum, weight decay (WD), and EDR (Exponential Decay Rate) range from 0.8 to 0.99, while the warmup epochs ($epoch_{warmup}$) are maintained within 10.

B. Dataset and Evaluation Metrics

1) Dataset construction and preprocessing: To comprehensively evaluate the detection performance of the proposed improved algorithm under complex scenarios, the publicly available LWDCD2020 dataset is adopted as the primary experimental dataset (as shown in Fig. 9). This dataset contains three common wheat disease categories—crown and root rot, leaf rust, and loose smut—as well as a healthy class, comprising a total of 12,000 images. Notably, the LWDCD2020 dataset is collected from diverse sources, integrating both field-acquired images and web-crawled data. Consequently, the images exhibit substantial variability in illumination conditions, viewing angles, background complexity, and object scale, enabling effective coverage of a wide range of real-world agricultural scenarios. This diversity endows the dataset with high complexity and representativeness, making it well-suited for evaluating the robustness and generalization capability of detection algorithms. It is worth noting that approximately 20% of the images are obtained via web crawling, which inevitably introduces inconsistencies between images and their annotations. These issues mainly include non-authentic wheat images, mislabeled samples, and images irrelevant to wheat. To ensure data quality and the reliability of experimental results, a rigorous manual verification and cleaning process is conducted to remove contaminated samples, thereby establishing a reliable dataset for subsequent experiments.

Within the disease samples, wheat plants infected with crown and root rot typically exhibit progressive leaf yellowing followed by wilting, accompanied by browning at the stem base, with symptoms intensifying as the disease develops. In contrast, leaf rust is characterized by nearly circular orange-red lesions irregularly distributed on leaf surfaces; as the disease progresses, these lesions may penetrate the leaf tissue, while dark brown to black elliptical lesions appear on the abaxial side. For loose smut, infected wheat spikes are initially covered by a thin gray membrane, which eventually ruptures as the disease advances, releasing black مسحوق-like spores. Diseased spikes are noticeably shorter than healthy ones, and both the main stem and tillers may be affected. Based on these observations, each disease category exhibits distinctive visual characteristics, while significant inter-class differences are maintained. These properties provide effective discriminative cues for detection models and further demonstrate the practical value of applying object detection methods to such tasks.

To enhance experimental efficiency, a parameter parallel mechanism was introduced to accelerate the execution of the IFDA algorithm. This mechanism transforms the water flow population into Resilient Distributed Datasets adapted for the

Spark platform, thereby realizing the parallel computation of the fitness of each individual within the population. This method significantly improves the execution efficiency of the experiments, providing robust technical support for subsequent experimental verification [16].

2) Evaluation metrics: This experiment adopts common evaluation metrics in the field of object detection to quantify the model's performance, primarily including Precision (P), Recall (R), and mean Average Precision (mAP).

The calculation formulas for Precision (P) and Recall (R) are as follows:

$$P = \frac{TP}{TP + FP} \tag{28}$$

$$R = \frac{TP}{TP + FN} \tag{29}$$

Where TP (True Positive) represents the number of correctly detected positive samples, FP indicates the number of falsely detected negative samples, and FN denotes the number of missed positive samples.

To comprehensively evaluate the detection effectiveness, we employ mAP@0.5 and mAP@0.5:0.95 as the core metrics:

mAP@0.5: Represents the mean of the average precision across all categories when the Intersection over Union (IoU) threshold is set to 0.5.

mAP@0.5:0.95: Represents the mean of the average precision when the IoU threshold varies from 0.5 to 0.95 (with a step size of 0.05).

TABLE II. COMPARATIVE EXPERIMENT OF MULTIPLE HEURISTIC EMA-YOLOV9 HYPERPARAMETER OPTIMIZATION ALGORITHM

Algorithm	Best Fitness	Average Fitness	Fitness Standard Deviation	Ranking Based on Average Value
EGA	0.309	0.2824	0.0155	12
GWO	0.4509	0.4329	0.0117	3
MVO	0.4341	0.4130	0.0206	4
SSA	0.4242	0.4053	0.0216	7
SCA	0.4325	0.4122	0.0156	5
BBO	0.3809	0.3304	0.0425	11
JSO	0.4142	0.3940	0.0232	10
MBO	0.4365	0.4047	0.0211	8
BA	0.4251	0.4044	0.0146	9
CSO	0.4203	0.4070	0.0108	6
FDA	0.4574	0.4402	0.0132	2
IFDA	**0.46856**	**0.4486**	**0.0144**	**1**

The population size ($Size_{pop}$) for all heuristic algorithms is set to 20, and the maximum number of iterations ($MaxIter$) is also set to 20. The number of neighbors used in IFDA and FDA is specified as 8. The hyperparameters for the remaining optimization algorithms are all configured to their default settings. Each algorithm is run independently 5 times to calculate the statistics for its optimal value, mean, and standard deviation.

Based on the statistical data analysis in Table II, although all tested algorithms exhibit a certain degree of inherent randomness, and no significant statistical differences are observed in the overall randomness distribution, a horizontal comparison of the baseline algorithms reveals that the Genetic Algorithm (GA) has the relatively lowest optimization effectiveness, whereas the Flow Direction Algorithm (FDA) demonstrates the best comprehensive performance. This empirical result strongly supports the assertion that FDA is the most suitable baseline optimization algorithm for the current EMA-YOLOv9 parameter optimization task.

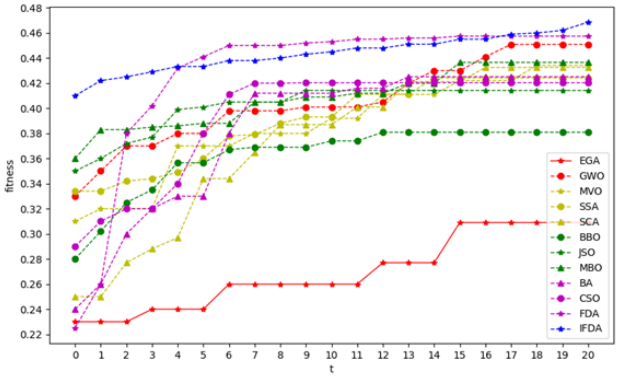


Fig. 10. Optimal fitness curve.

Fig. 10 illustrates the fitness convergence trajectories of the optimal results for each heuristic algorithm. It can be observed that the convergence curve representing IFDA exhibits a significant and steep upward trend during the initial iterations, indicating that, compared to other algorithms, IFDA possesses superior early convergence capability and optimization speed. This performance enhancement is primarily attributed to the improvement of the initialization solution strategy, which effectively elevates the quality of the initial population. Furthermore, the IFDA curve maintains a continuous upward trend in the later stages of iteration, indicating that the algorithm possesses a robust ability to escape local optima traps; the acquisition of this critical characteristic benefits from the introduction and optimization of the Lévy flight mechanism, thereby significantly enhancing the algorithm's global exploitation potential during the later stages of the search.

TABLE III. RESULTS OF ABLATION EXPERIMENT

Algorithm	Optimal fitness	Average adaptability	Accuracy	Recall rate	mAP_0.5
FDA	0.4574	0.4402	95.32	94.55	95.74
Initialization + FDA	0.4552	0.4427	95.10	94.33	95.52
Lévy flight mechanism + FDA	0.4662	0.4468	96.23	95.41	96.62
Initialization + Lévy flight mechanism + FDA	0.4686	0.4486	96.41	95.63	96.86

Specifically, after fine-tuning via the IFDA optimization strategy, the model achieved a significant increase of 5.09% in Precision, 5.11% in Recall, and 5.17% in mAP@0.5. These results fully substantiate the effectiveness of this optimization method in enhancing detection performance for complex object detection tasks.

C. Ablation Study

1) Ablation study of modules: To evaluate the effectiveness of the various improvement strategies within IFDA, an ablation study was conducted in this research. The YOLOv9 architecture and parameter settings remained consistent with those previously described, and the optimal parameters obtained through the optimization of each algorithm were utilized for training. The experimental results are presented in Table III: By combining the initialization strategy, the Lévy flight mechanism, and FDA, the model achieved the highest values in Precision, Recall, and mAP@0.5, confirming the effectiveness of the proposed improvements. Among them, the integration of Lévy flight significantly enhanced the performance of FDA, with the three metrics increasing by 0.95%, 0.91%, and 0.92%, respectively, compared to the original results; the overall performance improved by 0.13%, 0.13%, and 0.14%, respectively, compared to the original algorithm. Furthermore, the results of the ablation study regarding the various improvements of EMA-YOLOv9 are shown in Table IV.

TABLE IV. RESULTS OF ABLATION EXPERIMENT

YOLOv9	EMA	ResNet	IClou	**Accuracy %**	**Recall %**	**mAP_0.5 %**
√	×	×	×	90.80	90.05	90.77
√	√	×	×	91.21	90.26	91.48
√	×	√	×	90.92	90.11	91.29
√	×	×	√	90.98	90.16	91.33

The table systematically presents the performance of the YOLOv9-based object detection model under different strategy combinations, including the incorporation of EMA, ResNet backbone, and the ICIoU loss function, and their effects on Accuracy, Recall, and mAP\@0.5. The baseline YOLOv9 model achieves an Accuracy of 90.80%, Recall of 90.05%, and mAP of 90.77%. Incorporating EMA increases Accuracy to 91.21% and mAP to 91.48%, demonstrating its advantage in training stability and model generalization, with a limited improvement in Recall. Using the ResNet backbone alone results in an Accuracy of 90.92%, Recall of 90.11%, and mAP of 91.29%, indicating that enhanced feature extraction improves both localization and classification performance. Introducing only the ICIoU loss yields an Accuracy of 90.98%, Recall of 90.16%, and mAP of 91.33%, showing that the improved bounding box regression loss significantly optimizes detection precision. When all three strategies are applied simultaneously, the model achieves its highest performance with an Accuracy of 91.74%, Recall of 90.98%, and mAP of 92.09%, indicating a synergistic effect of feature enhancement, training stabilization, and bounding box optimization, thereby achieving overall optimal performance in both classification accuracy and object detection precision and validating the effectiveness of these strategies in enhancing YOLOv9.

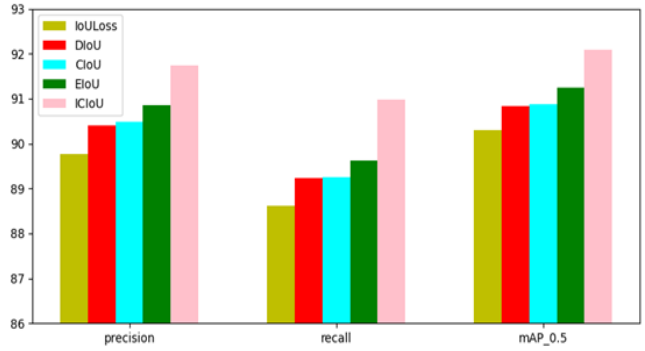


Fig. 11. Comparison of loss function experimental results.

2) Ablation study of loss functions: To verify the effectiveness of the proposed ICIoU loss function in the task of wheat disease and pest identification, we conducted a series of detailed comparative experiments between ICIoU and several other loss functions, including IoU loss, DIoU loss [17], CIoU loss, and EIoU loss [12]. The relevant experimental results are illustrated in Fig. 11. Compared to the baseline IoU loss function, the adoption of the ICIoU loss function improved the model's Precision, Recall, and mAP@0.5 by 2.19%, 2.66%, and 1.99%, respectively. Furthermore, compared to the CIoU loss function, the ICIoU loss function achieved gains of 1.39%, 1.94%, and 1.34% in Precision, Recall, and mAP@0.5, respectively.

Given the No Free Lunch theorem of stochastic optimization, it is impossible to directly determine a specific algorithm universally suited for the hyperparameter optimization of EMA-YOLOv9 [18]. Therefore, this paper compares the proposed IFDA with other heuristic algorithms to validate the advantages of FDA and IFDA in solving the hyperparameter optimization problem for EMA-YOLOv9, as shown in Table V. The heuristic algorithms included in this experiment are: Genetic Algorithm (GA), Grey Wolf Optimizer (GWO) [19], Multi-Verse Optimizer (MVO) [20], Salp Swarm Algorithm (SSA) [21], Sine Cosine Algorithm (SCA) [22], Biogeography-Based Optimization (BBO) [23], Jellyfish Search Optimizer (JSO) [24], Monarch Butterfly Optimization (MBO) [25], Bat Algorithm (BA) [26], and Cockroach Swarm Optimization (CSO) [27].

TABLE V. COMPARATIVE EXPERIMENTAL RESULTS

Method	**Precision %**	**Recall %**	**mAP_0.5 %**
Faster R-CNN[28]	86.73	86.79	87.60
YOLOv5+attention[29]	88.32	87.90	88.73
YOLO v8s+EIOU[30]	92.20	87.80	93.30
YOLOv8+ ShuffleNetV2+Senet[31]	88.90	81.40	88.20
GA-YOLOv5[7]	90.09	89.52	90.43
ABC-YOLOv5[9]	90.29	89.81	90.79
Ours	**96.41**	**95.63**	**96.86**

D. Comparison with SOTA Methods

To verify the advancement of IFDA-EMA-YOLOv9, we compared it with six State-of-the-Art (SOTA) methods. The experimental results are presented in *V*. The selected baseline models are as follows:

Faster R-CNN: Utilizes Faster R-CNN to detect wheat diseases and pests.

YOLOv5+attention: Optimizes YOLOv5 using an improved attention sub-module, and combines BiFPN with fast normalization fusion to weight the fused features.

YOLOv8s+EIOU: Improves YOLOv8s by employing efficient RepGFPN feature fusion and the EIoU loss function.

YOLOv8+ShuffleNet V2+Senet: Employs the lightweight convolutional neural network ShuffleNet V2 as the backbone network of YOLOv8, improving it by adding a small object detection layer and the Senet attention mechanism.

GA-YOLOv5: Represents the optimization of YOLOv5 hyperparameters using its built-in Genetic Algorithm (GA).

ABC-YOLOv5: Represents the optimization of YOLOv5 hyperparameters using the Artificial Bee Colony (ABC) algorithm.

The proposed method demonstrates superior performance compared to GA-YOLOv5 and ABC-YOLOv5, thereby establishing the significant advantages of the IFDA method over traditional optimization algorithms such as the Genetic Algorithm (GA) and Artificial Bee Colony (ABC). Specifically, on the publicly available LWDCD2020 dataset, compared to the best existing method excluding EMA-YOLOv9, the proposed method increases Precision by 6.4% and Recall by 5.94%. Furthermore, the core metric mAP@0.5 is improved by 6.66%. These results indicate that the methodology proposed in this paper has achieved significant progress in related fields. Fig. 12 illustrates the visualized detection results of IFDA-EMA-YOLOv9 across various categories of samples.



(a) Nutritious wheat (b) Crown rot and root rot

(c) leaf rust disease (d) Loose Smut of Wheat

Fig. 12. Visual display of detection result.

VI. Conclusion

This paper proposes a wheat disease detection method based on the improved EMA-YOLOv9 and IFDA. This method integrates multi-scale attention and residual connections to enhance feature extraction, introduces an auxiliary box to optimize localization accuracy, and utilizes the IFDA algorithm—which combines chaotic maps and Lévy flight—to achieve adaptive hyperparameter optimization. Experiments demonstrate that, on the LWDCD2020 dataset, the proposed method improves Precision, Recall, and mAP@0.5 by 6.4%, 5.94%, and 6.66%, respectively, compared to SOTA methods. Additionally, under the experimental environment considered in this study, IFDA-EMA-YOLOv9 achieves a detection frame rate (FPS) of approximately 29.8, which is sufficient to meet the requirements of wheat disease detection tasks. Future work will focus on model lightweighting, exploring its real-time deployment on mobile devices, and its generalized application across multiple scenarios.

References

[1] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in Proc. European Conference on Computer Vision, Cham: Springer Nature Switzerland, 2024.

[2] Li, B., & Fan, J.-W. (2025). Deep Learning-oriented Literature Review and Experimental Analysis of Ship Target Detection. Science Technology and Engineering, 25(27), 11427–11450. https://doi.org/10.12404/j.issn.1671-1815.2409633

[3] Yue, Y. J., et al. "Object detection of tomato fruit based on cascade RCNN." J. Science Technology and Engineering 21 (2021): 2387.

[4] Qiao, S., Yuan, Y., & Qi, R. (2025). Meter-YOLOv8n: A Lightweight and Efficient Algorithm for Word-Wheel Water Meter Reading Recognition. International Journal of Advanced Computer Science & Applications, 16(4).

[5] H. Zhang, C. Xu, and S. Zhang, "Inner-IoU: more effective intersection over union loss with auxiliary bounding box," arXiv preprint arXiv:2311.02877, 2023.

[6] S. Ma and Y. Xu, "Mpdiou: a loss for efficient and accurate bounding box regression," arXiv preprint arXiv:2307.07662, 2023.

[7] L. Popek, R. Perz, G. Galinski, and A. Abratanski, "Optimization of animal detection in thermal images using YOLO architecture," Int. J. Electron. Telecommun., vol. 69, no. 4, pp. 693–700, 2023.

[8] H. Mahmudah, A. S. Aisjah, S. Arifin, and C. A. Prastyantyo, "iYOLOV7-TPE-SS: Leveraging improved YOLO model with multilevel hyperparameter optimization for road damage detection on edge devices," IEEE Access, vol. 13, pp. 1–15, 2025.

[9] A. Araman, I. Pacal, A. Basturk, et al., "Robust real-time polyp detection system design based on YOLO algorithms by optimizing activation functions and hyper-parameters with artificial bee colony (ABC)," Expert Syst. Appl., vol. 221, p. 119741, 2023.

[10] Yang, Q.-H., Yang, G.-C., & Zhong, S.-H. (2025). Hyperparameter Optimization of YOLO Model Based on Orthogonal Optimization Strategy. Science Technology and Engineering, 25(4), 1573–1579. https://doi.org/10.12404/j.issn.1671-1815.2309596

[11] J. F. O'Callaghan and D. M. Mark, "The extraction of drainage networks from digital elevation data," Comput. Vis. Graph. Image Process., vol. 28, no. 3, pp. 323–344, 1984.

[12] H. Arami, M. V. Anaraki, S. Farzin, and S. Mirjalili, "Flow direction algorithm (FDA): a novel optimization approach for solving optimization problems," Comput. Ind. Eng., vol. 156, p. 107224, 2021.

[13] D. Uyang, et al., "Efficient multi-scale attention module with cross-spatial learning," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2023, pp. 1–5.

[14] Y. F. Hang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," Neurocomputing, vol. 506, pp. 146–157, 2022.

[15] H. Lin and C. Wang, "DIGWO-N-BEATS: An evolutionary time series prediction method for situation prediction," Inf. Sci., vol. 664, p. 120316, 2024.

[16] H. Lin, C. Wang, and Q. Hao, "A novel personality detection method based on high-dimensional psycholinguistic features and improved distributed gray wolf optimizer for feature selection," Inf. Process. Manage., vol. 60, no. 2, p. 103217, 2023.

[17] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, 2020, pp. 12993–13000.

[18] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," IEEE Trans. Evol. Comput., vol. 1, no. 1, pp. 67–82, 1997.

[19] S. N. Makhadmeh et al., "Recent Advances in Grey Wolf Optimizer, its Versions and Applications: Review," in IEEE Access, vol. 12, pp. 22991-23028, 2024, doi: 10.1109/ACCESS.2023.3304889.

[20] L. Abualigah, "Multi-verse optimizer algorithm: a comprehensive survey of its results, variants, and applications," Neural Comput. & Applic., vol. 32, pp. 12381–12401, 2020, doi: 10.1007/s00521-020-04839-1.

[21] L. Abualigah, M. Shehab, M. Alshinwan, and H. Alabool, "Salp swarm algorithm: A comprehensive survey," Neural Comput. Appl., vol. 32, no. 15, pp. 11195–11215, 2020.

[22] Q. S. Hamad, S. A. M. Saleh, S. A. Suandi, et al., "A Review of Enhancing Sine Cosine Algorithm: Common Approaches for Improved Metaheuristic Algorithms," Arch. Computat. Methods Eng., vol. 32, pp. 2549–2606, 2025, doi: 10.1007/s11831-024-10218-z.

[23] W. Guo, M. Chen, L. Wang, et al., "A survey of biogeography-based optimization," Neural Comput. & Applic., vol. 28, pp. 1909–1926, 2017, doi: 10.1007/s00521-016-2179-x.

[24] J. S. Hou and A. Molla, "Recent advances in use of bio-inspired jellyfish search algorithm for solving optimization problems," Sci. Rep., vol. 12, no. 1, p. 19157, 2022.

[25] Y. Feng, S. Deb, G. Wang, A. H. Alavi, "Monarch butterfly optimization: A comprehensive review," Expert Systems with Applications, vol. 168, pp. 114418, 2021, doi: 10.1016/j.eswa.2020.114418.

[26] T. Agarwal, V. Kumar, "A Systematic Review on Bat Algorithm: Theoretical Foundation, Variants, and Applications," Arch. Computat. Methods Eng., vol. 29, pp. 2707–2736, 2022, doi: 10.1007/s11831-021-09673-9.

[27] L. Cheng, et al., "A Review on Cockroach Swarm Optimization Algorithm," in Proceedings of the 12th International Conference on Computer Engineering and Networks, CENet 2022, vol. 961, Springer, Singapore, 2022, doi: 10.1007/978-981-19-6901-0_151.

[28] B. S. Bari, M. N. Islam, M. Rashid, M. J. Hasan, M. A. Razman, R. M. Musa, and A. P. A. Majeed, "A real-time approach of diagnosing rice leaf disease using deep learning-based Faster R-CNN framework," PeerJ Comput. Sci., vol. 7, p. e432, 2021.

[29] H. Wang, S. Shang, D. Wang, X. He, K. Feng, and H. Zhu, "Plant disease detection and classification method based on the optimized lightweight YOLOv5 model," Agriculture, vol. 12, no. 7, p. 931, 2022.

[30] Trinh, D. C., Mac, A. T., Dang, K. G., Nguyen, H. T., Nguyen, H. T., & Bui, T. D. (2024). Alpha-EIOU-YOLOv8: an improved algorithm for rice leaf disease detection. AgriEngineering, 6(1), 302-317.

[31] Cen, X., Lu, S., & Qian, T. (2025). YOLO-LCE: A Lightweight YOLOv8 Model for Agricultural Pest Detection. Agronomy, 15(9), 2022

Energy-Efficient Multi-Hop LoRa Communication in Forested Environments via Proximal Policy Optimization

Muhd Kahfi bin Jumali[1], Lim Kit Guan[2]*, Ervin Gubin Moung[3],
Lorita Angeline[4], Tianlei Wang[5], Kenneth Teo Tze Kin[6]*
Modelling, Simulation & Computing Laboratory-Faculty of Engineering,
Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia[1, 2, 4, 6]
Data Technology and Applications Research Group-Faculty of Computing and Informatics,
Universiti Malaysia Sabah, Kota Kinabalu, Malaysia[3]
Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen, China[5]

***Abstract*—Multi-hop LoRa networks extend coverage for large-scale Internet of Things deployments but are severely limited by interference-induced collisions, retransmissions, and rapid battery depletion of relay nodes. Conventional routing strategies that minimize hop count or rely on static heuristics fail to account for dynamic medium contention and its impact on energy consumption and reliability. This paper proposes a Proximal Policy Optimization (PPO)–based routing framework for multi-hop LoRa networks that learns interference-aware and energy-efficient routing policies through reinforcement learning. A discrete-event simulation framework is developed to model LoRa physical-layer behaviour, co-spreading-factor interference, adaptive data rate control, and battery-limited relay nodes under multi-source traffic. The routing problem is formulated as a Markov Decision Process (MDP) in which the PPO agent selects next-hop relays based on local topology, relay load, and channel occupancy, while physical-layer parameters are adapted independently using a standards-inspired physical-layer parameters are adapted independently using a standards-inspired ADR mechanism (ADR) mechanism. Simulation results show that the proposed approach achieves a packet delivery ratio of up to 73.7%, reduces collision rates by approximately 46% compared with Random routing, and lowers the average energy consumption per delivered packet to about 206 mJ, outperforming Shortest Path and Ad hoc On-Demand Distance Vector (AODV)-like routing. These gains are achieved by learning spatially diverse routing paths that mitigate relay congestion and reduce collision-induced retransmissions.**

***Keywords*—LoRa; Multi-Hop; PPO; ADR; MDP-Markov Decision Process; Reinforcement-Learning; PDR**

I. Introduction

The rapid growth of Internet of Things (IoT) [1] applications in smart agriculture, environmental monitoring, and industrial automation has accelerated the adoption of Low Power Wide Area Networks (LPWANs) as a cost-effective solution for long-range, energy-constrained wireless connectivity [2]. Among LPWAN technologies, Long Range (LoRa) has emerged as a particularly attractive option due to its ability to provide kilometer-scale coverage while enabling multi-year battery operation [3]. However, the star topology employed in Long Range Wide Area Network (LoRaWAN), where end devices communicate directly with a gateway, becomes ineffective in scenarios where distance, terrain, or urban obstructions prevent reliable direct links, as empirically demonstrated in forested and vegetated deployments [4], [5], [6]. To overcome this limitation, multi-hop LoRa networks introduce relay-assisted communication, extending coverage while preserving LoRa's low-power characteristics [7].

Despite their potential, multi-hop LoRa networks face significant routing challenges under multi-source traffic. Unlike traditional wireless sensor networks with centralized scheduling or orthogonal channel access, LoRa operates under pure Additive Links On-line Hawaii Area (ALOHA) and strict duty-cycle constraints [8], limiting coordination. When multiple sources independently select relays based on local metrics such as distance or link quality, traffic converges on a small subset of relays, creating persistent collision hotspots. These collisions increase retransmission, waste energy through long transmissions, and accelerate relay battery depletion. Although Adaptive Data Rate (ADR) improves link robustness, it does not mitigate contention among competing sources, leaving collision-aware routing an open challenge.

Existing routing approaches for multi-hop LoRa [9] networks fall into three broad categories. Distance-based heuristics, such as shortest path routing, offer simplicity but suffer from severe load imbalance and collision concentration. Adaptive protocols inspired by Ad hoc On-Demand Distance Vector (AODV) incorporate energy or load-aware metrics to improve fairness yet still rely on static decision rules that are insufficient under dynamic multi-source contention. More recent studies have applied reinforcement learning to LoRa networks, primarily focusing on physical-layer parameter optimization such as spreading factor selection [10]. While effective at the link level, these approaches do not address network-layer routing coordination across competing sources, leading to suboptimal performance in multi-hop scenarios.

Energy efficiency is a central concern in this context, as relay nodes are typically battery-powered and must forward traffic from multiple sources [11]. In interference-limited LoRa networks, energy consumption is dominated not by hop count

*Corresponding author.

alone but by collision-induced retransmissions and the use of high spreading factors on congested links. Consequently, routing strategies that ignore multi-source interaction often incur excessive energy expenditure and uneven battery depletion, directly reducing network lifetime [12]. The fundamental challenge is therefore to design routing mechanisms that jointly improve packet delivery reliability, minimize energy consumption, and balance relay utilization under dynamic traffic and interference conditions.

Recent advances in deep reinforcement learning [13] have shown strong potential for addressing such multi-objective optimization problems in complex networked systems. Proximal Policy Optimization (PPO) offers stable and sample-efficient policy learning in non-stationary environments, making it well suited for multi-source routing scenarios where traffic patterns and interference evolve over time [14]. When provided with appropriate observability into relay utilization and channel conditions, PPO can learn adaptive routing behaviors without requiring explicit coordination or message exchange between sources.

Motivated by this observation, this paper proposes a PPO-based routing framework for multi-hop LoRa networks that explicitly accounts for multi-source interaction. By augmenting the routing state representation with features capturing relay usage frequency and temporal recency, the proposed approach enables implicit coordination among sources, allowing routing decisions to proactively avoid congested relays. Physical-layer parameters are adapted independently using a standards-inspired ADR mechanism [15], ensuring a clear separation between routing optimization and link-level adaptation.

The main contributions of this work can be summarized as follows:

- Multi-source–aware state features are introduced to enable routing agents to detect and avoid relays that are currently utilized by competing sources, thereby reducing collision probability without incurring explicit coordination overhead.
- An energy and reliability-aware reward function is designed to balance packet delivery, energy consumption, and collision avoidance in battery-limited relay networks.
- A curriculum-based PPO training strategy is developed to accelerate convergence toward collision-aware routing policies.
- Finally, through comprehensive simulation-based evaluation, it is demonstrated that the proposed framework achieves improved packet delivery reliability, reduced collision rates, and enhanced energy efficiency compared with conventional routing approaches, while exhibiting emergent spatial path separation that mitigates relay energy hotspots.

The remainder of this paper is organized as follows. Section II reviews related work on multi-hop LoRa routing and reinforcement learning for wireless networks. Section III presents the system model, including network architecture, collision dynamics, and energy consumption. Section IV details the proposed PPO-based routing framework. Section V discusses performance evaluation and key insights. Finally, conclusions are drawn in Section VI.

II. Related Work

LPWAN [16] technologies were developed to provide long-range connectivity under stringent energy constraints, making them suitable for large-scale IoT deployments [17] where frequent battery replacement is infeasible. Comparative surveys position LoRa/LoRaWAN among the most widely adopted LPWAN options because it provides kilometers-scale coverage with low device complexity and low power operation, albeit with trade-offs in capacity and interference robustness under ALOHA-like access. In practical deployments, coverage and reliability are highly scenario-dependent (terrain, foliage, obstacles, antenna height), motivating designs that go beyond conventional single-hop LoRaWAN star topologies when direct gateway connectivity is unreliable.

A major limitation of LoRa networks especially under multi-node contention is that performance can degrade rapidly due to collisions, imperfect orthogonality among spreading factors, and capture effects [18]. Early experimental and analytical studies show that "scaling" LoRa is non-trivial because concurrent transmissions can interfere even when nominally using different Spreading Factor (SF), while co-SF collisions depend on relative received powers and overlap in time [19]. Complementary simulation-based investigations further quantify how packet delivery probability is impacted by offered load, SF distribution, and channel contention, reinforcing that interference modelling choices (co-SF and inter-SF rejection assumptions, capture threshold, traffic model) materially affect energy and reliability conclusions.

In multi-hop settings, these issues can be amplified: relays introduce additional transmissions (hence additional channel occupancy), and traffic tends to concentrate around a subset of "good" relays (those that make strong progress toward the gateway), increasing the probability of repeated collisions and retransmission-driven energy waste [20]. This motivates routing policies that are explicitly collision-aware and load-aware rather than purely distance-optimal. LoRaWAN's ADR [21] is a standardized mechanism designed primarily for star topologies: the network server adjusts SF and transmits power using link-quality statistics (commonly using margin-based logic over recent Signal-to-Noise Ratio (SNR) / Received Signal Strength Indicator (RSSI) history) [22]. ADR is widely used because reducing SF and/or transmit power can substantially reduce time-on-air and energy per message when link margin is high, while increasing SF/power improves robustness when the link is marginal.

However, ADR alone does not resolve contention externalities: even when each node's physical settings are individually reasonable, collisions can still dominate under shared-medium access particularly when multiple nodes select similar SFs and transmit times. Thus, while ADR addresses the "link adaptation" dimension, it typically does not optimize network-layer forwarding decisions or coordinate multi-source relay usage in multi-hop mesh behaviours [23].

Multi-hop and mesh extensions [24] for LoRa have been explored to address coverage holes and obstacles, proposing lightweight routing protocols (often distance-vector-like, flooding-assisted discovery, or heuristic metrics such as hop count, link quality, or airtime cost). A recent comprehensive survey consolidates these directions, highlighting that many LoRa mesh proposals prioritize simplicity and feasibility over strong optimality guarantees, and often assume limited coordination under duty-cycle constraints. Empirical protocol studies also demonstrate that practical mesh routing must manage not only reachability and hop efficiency but also traffic concentration and fairness, because a small number of relays can become persistent bottlenecks [25].

Importantly, classical ad hoc routing protocols such as AODV were developed for higher-rate radios and different interference regimes, relying on route discovery and maintenance logic that may not directly translate to LoRa's long airtime, low duty cycle, and collision-dominated shared channel. Nevertheless, AODV remains a common baseline because it captures the value of adaptive path selection versus static shortest-path heuristics. Learning-based methods have been applied primarily to LoRaWAN parameter adaptation (e.g., selecting SF / Transmission Power (TP) / channel assignments to improve energy efficiency and reliability). Recent Deep Reinforcement Learning (DRL) work shows that intelligent control can outperform fixed heuristics in dynamic environments, especially when objective functions include energy and performance constraints [26]. In parallel, PPO is widely recognized as a stable policy-gradient algorithm for complex decision-making problems, due to its clipped surrogate objective that reduce destructive policy updates [27].

Despite this progress, most learning-driven LoRa papers focus on star topologies and Physical Layer (PHY) parameter selection rather than joint routing/forwarding under multi-hop mesh contention [28]. In a multi-source, battery-limited relay mesh, the key difficulty is not only choosing energy-efficient PHY settings but also preventing multiple sources from repeatedly converging on the same high-quality relay-creating collision hotspots and unfair battery depletion. Across the above literature, a consistent gap is the lack of explicit multi-source awareness in routing decisions for multi-hop LoRa meshes [29]. Existing mesh routing protocols and classical baselines typically optimize progress metrics (distance/hops/link quality) but do not directly encode "competition" from other sources (who is using which relays and how recently), even though that competition is the driver of collision synchronization and relay hotspot depletion in shared-medium LoRa. At the same time, ADR-based PHY adaptation optimizes link margin but does not solve contention-driven losses or relay fairness.

From the above, a clear gap emerges where LoRa scalability or interference studies characterize collision behaviour and airtime-driven limitations but do not provide a network-layer mechanism for multi-source relay competition in battery-limited multi-hop settings. Multi-hop LoRa routing protocols propose metrics such as Time on Air (ToA)-aware distance-vector routing and cross-layer scheduling but typically lack explicit multi-source awareness signals that help avoid relay hotspots formed by concurrent sources selecting the same "best" relays.

III. Methodology

A. Network Architecture

The considered multi-hop LoRa network is represented as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V}$ comprises source nodes, battery-powered relay nodes, and a gateway acting as the final data sink. A directed link $(i, j) \in \mathcal{E}$ exists if node j lies within the communication range of node i. Each source periodically generates uplink packets that must be forwarded through one or more relay hops to reach the gateway. Relay nodes operate in a store-and-forward manner and are constrained by finite energy budgets, while the gateway is assumed to have unlimited energy resources. The network topology remains static during each simulation run, whereas wireless channel conditions, interference patterns, and medium access timing evolve dynamically. LoRa Mesh Network Topology is shown in Fig. 1.

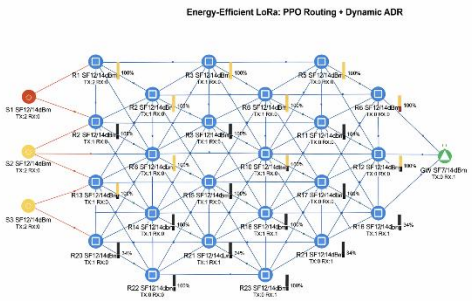


Fig. 1. LoRa Mesh Network Topology

B. Energy Consumption Model

The energy consumption of each relay node is determined by the power used for transmission and the corresponding LoRa time-on-air. Specifically, the energy consumed for transmission from node i to node j is expressed as

$$E_{i,j}^{\text{tx}} = I_{\text{tx}}(P_i^{\text{tx}})\, V\, T_{\text{air}} \tag{1}$$

where $P_{i,j}^{\text{tx}}$ denotes the transmit power selected for the link and $T_{\text{air}}(SF_{i,j})$ represents the airtime associated with the selected spreading factor. The airtime consists of preamble and payload components and increases exponentially with the spreading factor due to the reduced bit rate. Consequently, higher spreading factors result in significantly longer transmission durations and increased energy expenditure. Each relay node maintains a residual energy state that is updated after every transmission according to:

$$E_i^{\text{res}}(t+1) = E_i^{\text{res}}(t) - \sum E_{i,j}^{\text{tx}} \tag{2}$$

Equation (2) updates the residual energy of relay i after time step t by subtracting the total transmission energy spent on all forwarding events executed by that relay during the step. Here, $E_i^{\text{res}}(t)$ denotes the remaining battery energy at time t, and $E_{i,j}^{\text{tx}}$ represents the energy consumed when node i transmits a packet to next hop j. In our simulator, $E_{i,j}^{\text{tx}}$ is computed from the radio's current draw at the selected transmit power and the packet time-on-air, i.e., $E_{i,j}^{\text{tx}} = I_{\text{tx}}(P_{\text{tx}})\, V\, T_{\text{OA}}$, ensuring that higher spreading factors or power levels incur higher energy.

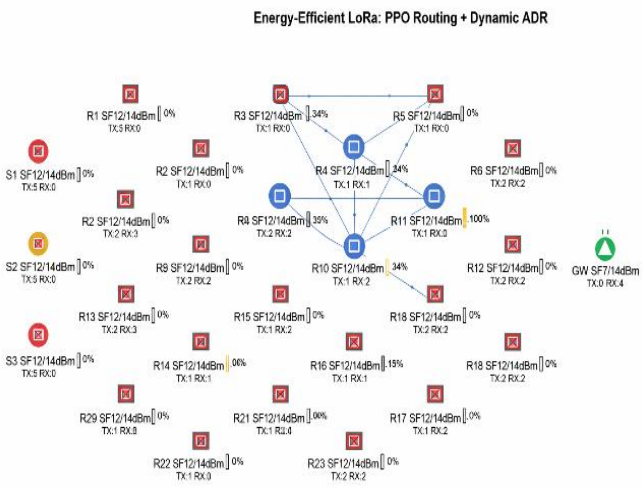


Fig. 2. Battery depletion.

Costs through longer airtime and larger current consumption. Once $E_i^{\text{res}}(t) \le 0$, relay i is treated as depleted (dead) and is removed from candidate next-hop sets, which can force route reconfiguration and directly impacts network lifetime. Fig.2 shows the Battery depletion.

C. Signal Propagation Model

Wireless signal propagation is modelled using a forest path-loss [30] formulation augmented with log-normal shadow fading to capture the effects of vegetation and terrain irregularities. To capture realistic rural/forested deployments, the wireless channel is modelled using a log-distance path-loss formulation with an additional vegetation attenuation term and log-normal shadow fading, consistent with empirical LoRaWAN path-loss characterizations in forested and vegetated environments reported in [5], [6], [31], [32]. The vegetation attenuation term κd represents the specific attenuation through foliage as standardized in ITU-R Recommendation P.833-10 [33]. For a transmitter–receiver separation distance $d_{i,j}$ (in meters), the path loss is expressed as:

$$PL(d) = PL(1\text{ m}) + 10n\log_{10}(d) + \kappa d + X_\sigma \quad (3)$$

where $PL(1\text{ m})$ denotes the free-space path loss at 1 meter, n is the path-loss exponent, κ is a vegetation attenuation coefficient (in dB/m), and $X_\sigma \sim \mathcal{N}(0,\sigma^2)$models log-normal shadow fading. In our implementation, $PL(1\text{ m})$ is obtained from the standard free-space expression at carrier frequency f,

$$RSSI_{i,j} = P_i^{\text{tx}} + G_t + G_r - PL(d_{i,j}) \quad (4)$$

where P_i^{tx} is the transmit power (dBm) and G_t and G_r are antenna gains (dBi). This formulation directly matches the simulator's computation $RSSI = P^{tx} + 2G - PL$.

$$SNR_{i,j} = RSSI_{i,j} - N_0 \quad (5)$$

$$N_0 = -174 + 10\log_{10}(BW) + NF \quad (6)$$

with N_0 denoting the noise floor. Successful packet reception requires both the RSSI and SNR to exceed the minimum demodulation thresholds determined by the selected spreading factor [34]. If either condition is violated, the packet is considered lost due to weak signal conditions [35].Demodulation Criteria in Table I

TABLE I. DEMODULATION CRITERIA

SF	Receiver Sensitivity Threshold (dBm)	Min SNR (dB)
7	-123	-7.5
8	-126	-10
9	-129	-12.5
10	-132	-15
11	-134.5	-17.5
12	-137	-20

Transmission duration is modeled using the LoRa time-on-air expression, which directly influences both channel occupancy (collision probability) and energy consumption. The symbol duration is

$$T_{\text{sym}} = \frac{2^{SF}}{BW} \quad (7)$$

$$T_{\text{preamble}} = (N_{\text{preamble}} + 4.25)T_{\text{sym}} \quad (8)$$

$$Tpayload = Npayload_symb \cdot Tsym \quad (9)$$

$$N_{\text{payload}} = 8 + \max\left(\left\lceil \frac{8PL - 4SF + 28 + 16CRC - 20IH}{4(SF - 2DE)} \right\rceil (CR + 4), 0\right) \quad (10)$$

where N_{preamble} is the programmed preamble length. The payload duration depends on the number of payload symbols N_{payload} and is payload length (bytes), $CRC \in 0,1$ indicates CRC enablement, $IH \in 0,1$ indicates implicit header mode, and $DE \in 0,1$ is the low-data-rate optimization flag (typically enabled for $SF \ge 11$). This expression is consistent with the simulator's airtime computation used for every transmission event.

D. Co-SF interference & Collision Decision

Collisions [36] are modeled using co-spreading-factor overlaps, consistent with the simulator's event-based tracking of concurrent transmissions at each SF. During the reception of packet p transmitted from node i to node j, the set of overlapping interferers using the same SF is denoted $\mathcal{J}_j^{(SF)}$. Interferer received powers are accumulated in the linear domain:

$$P_{k,j}^{(\text{lin})} = 10^{\frac{RSSI_{k,j}}{10}} \quad (11)$$

where $RSSI_{k,j}$ is expressed in dBm. The aggregate interference power at node j is then obtained by summing the linear powers of all interfering transmitters,

$$I_j = \sum_{k \in \mathcal{J}_j^{(SF)} \setminus i} P_{k,j}^{(\text{lin})} \quad (12)$$

The signal-to-interference ratio at the receiver is computed as:

$$SIR_{i,j} = 10\log_{10}\left(\frac{10^{RSSI_{i,j}/10}}{I_j + \epsilon}\right) \quad (13)$$

where ϵ is a small constant introduced to avoid numerical instability when no concurrent interferers are present. A packet

collision is declared when the computed Signal-to-Interference Ratio (SIR) falls below a predefined capture threshold γ_{th}

$$\text{Collision occurs if } SIR_{i,j} < \gamma_{\text{th}} \quad (14)$$

In this study, a conservative capture threshold is employed, reflecting the imperfect orthogonality of LoRa spreading factors under real-world synchronization and fading conditions. When a collision occurs, the packet is considered lost and removed from the reception buffer, triggering retransmission attempts and additional energy consumption. This collision model directly links medium access contention, routing decisions, and physical-layer parameter selection, thereby allowing the reinforcement learning agent to implicitly learn collision-aware routing behavior.

E. Dynamic Adaptive Data Rate (ADR) Mechanism

To adapt physical-layer parameters in response to time-varying channel conditions, a Dynamic ADR mechanism is employed for all routing schemes [37]. ADR operates independently of the routing algorithm and is applied uniformly to ensure a fair comparison between PPO-based routing and baseline methods. For each transmitter–receiver link (i, j), the ADR controller maintains a sliding window of recent received signal strength and signal-to-noise ratio measurements. Let $\mathcal{H}_{i,j} = \{SNR_k\}_{k=1}^{W}$ denote the most recent W SNR observations collected at the receiver. The average link margin is computed as:

$$M_{i,j} = \text{SNR}_{i,j} - \text{SNR}_{min}(\text{SF}) \quad (16)$$

where $\bar{SNR}_{i,j}$ is the mean SNR over the observation window and $SNR_{\min}(SF)$ is the minimum demodulation threshold for the currently selected spreading factor. When the link margin exceeds a predefined upper threshold, the ADR controller attempts to reduce energy consumption by decreasing the spreading factor or transmit power, prioritizing spreading factor reduction due to its exponential impact on time-on-air. Conversely, when the margin falls below a lower threshold, the controller increases the spreading factor or transmits power to restore link reliability. If neither adjustment is feasible due to parameter bounds, the current configuration is retained. This adaptive process balances robustness and energy efficiency while avoiding excessive packet loss. ADR updates are performed periodically and only after sufficient measurement samples are available, preventing unstable oscillations in parameter selection.

Algorithm 1 Dynamic Adaptive Data Rate (ADR) Control

```
INPUT: Node ID, Current SF, Current TX Power, Link Quality History
OUTPUT: Optimized SF and TX Power

FUNCTION UpdateLinkQuality (nodeID, RSSI, SNR):
   history_RSSI [nodeID]. append (RSSI)
   history_SNR [nodeID]. append (SNR)
   Keep only the most recent N measurements
END FUNCTION

CalculateADR (nodeID, SF_current, P_current):

   IF length (history_RSSI [nodeID]) < 10 THEN
      RETURN (SF_current, P_current)
   END IF

   RSSI_avg ← (1 / N) × Σ (i = 1 to N) RSSI_i
   SNR_avg  ← (1 / N) × Σ (i = 1 to N) SNR_i

   Margin_RSSI ← RSSI_avg – Sensitivity [SF_current]
   Margin_SNR  ← SNR_avg  − MinSNR [SF_current]

   SF_new ← SF_current
   P_new  ← P_current

   IF (Margin_RSSI > 15 dB) AND (Margin_SNR > 10 dB)
THEN
      IF SF_current > SF_min THEN
         SF_new ← SF_current − 1
      END IF
   ELSE IF (Margin_RSSI < 5 dB) OR (Margin_SNR < 2 dB)
THEN
      IF SF_current < SF_max THEN
         SF_new ← SF_current + 1
      END IF
   END IF

   Margin_RSSI′  ← RSSI_avg − Sensitivity [SF_new]

   IF (Margin_RSSI′  > 20 dB) AND (P_current > P_min)
THEN
      P_new ← max(P_min, P_current − 3 dB)
   ELSE IF (Margin_RSSI′  < 8 dB) AND (P_current <
P_max) THEN
      P_new ← min(P_max, P_current + 3 dB)
   END IF

   RETURN (SF_new, P_new)
```

IV. Proposed Learning Model

A. Problem Formulation

Routing in a multi-hop LoRa mesh network is inherently sequential and stochastic due to time-varying interference, shared medium access, and battery depletion of relay nodes. At each forwarding step, a relay must select a next-hop neighbor without full knowledge of future channel conditions or competing transmissions from other sources. This makes classical shortest-path or greedy routing suboptimal in the presence of contention and energy imbalance. To address this challenge, the routing problem is formulated as a Markov Decision Process (MDP) [38], enabling the use of reinforcement learning to derive an adaptive routing policy. The MDP is defined by the tuple (S, A, P, R, γ) where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ denotes the state transition dynamics, $\mathcal{R}$ is the reward function, and $\gamma \in (0,1]$ is the discount factor. The objective is to learn a routing policy $\pi_\theta(a \mid s)$, parameterized by θ, that maximizes the expected cumulative discounted reward over an episode, corresponding to energy-efficient and reliable end-to-end packet delivery.

1) State space design. The state observed by the PPO agent encodes both topological progress and medium access conditions relevant to routing decisions [39]. For each forwarding node, a fixed-length state vector is constructed by

aggregating features associated with up to K candidate next-hop relays. In this study, K = 10 is used to bound the action space. Let $\mathbf{s}_t \in \mathbb{R}^{16K}$ denote the state at decision step t. For each candidate relay k, the following normalized features are included: relative distance to the gateway, hop distance from the current node, normalized progress toward the gateway, residual energy ratio, historical load ratio, averaged RSSI and SNR statistics, channel occupancy at the candidate's spreading factor, node-level and link-level collision rates, local traffic density, spreading-factor congestion, transmission recency, and relay load indicators. To explicitly address multi-source contention, two additional features are incorporated: the number of other sources recently using the same relay and the recency of such usage. The state vector is expressed as

$$st = [f1, t, f2, t, \dots, fK, t], fk, t \in R16 \quad (17)$$

where zero-padding is applied when fewer than K candidates are available. This formulation ensures a fixed-dimensional input suitable for neural network training while preserving local routing context and contention awareness.

2) Action space & feasibility Masking. The action space corresponds to the selection of the next-hop relay among the candidate neighbors [40]. At each decision step t, the agent selects an action

$$at \in 1, 2, \dots, K, \quad (18)$$

where each index corresponds to a specific candidate relay in the ordered neighbor list. Actions associated with invalid or non-existent candidates are masked during policy evaluation to prevent infeasible selections. Importantly, the PPO agent optimizes routing decisions only; physical-layer parameters such as spreading factors and transmit power are adapted independently by the ADR controller.

The reward function is designed to jointly incentivize reliable packet delivery, energy efficiency, collision avoidance, and load balancing across relay nodes [41]. Let r_t denote the reward assigned after an episode completes. A large positive reward is granted when a packet is successfully delivered to the gateway, scaled by the current packet delivery ratio to encourage global reliability. Conversely, strong penalties are imposed for packet loss due to collisions or weak signal conditions, reflecting the high energy cost of retransmissions. Energy efficiency is explicitly encouraged by penalizing excessive transmission energy, while hop-count penalties discourage unnecessarily long routes. To mitigate relay hotspots and promote fairness, additional penalties are applied when routing paths overlap significantly with those of other concurrent sources or when relay load imbalance is detected. Formally, the reward can be expressed as:

$$r = R_s ucc - \alpha E - \beta H - \delta O, if delivered$$
$$r = -R_f ail - \alpha E - \beta H, otherwise \quad (19)$$

Here, $R_s ucc$, and $R_f ail$ are positive constants rewarding successful delivery and penalizing packet failure, respectively. The coefficients α, β, and δ control the relative importance of energy consumption, hop count, and path overlap. By penalizing overlapping relay usage, the reward function discourages routing decisions that create relay hotspots and persistent collisions. This formulation enables the agent to learn routing strategies that trade minimal hop count for reduced contention and energy waste when necessary.

B. Proximal Policy Optimization (PPO) Architecture

The routing policy is optimized using PPO [42], which improves training stability by constraining policy updates. PPO maximizes a clipped surrogate objective defined as

$$L^{PPO}(\theta) = E_t[\min(r_t(\theta) A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) A_t)] \quad (20)$$

where,

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)} \quad (21)$$

is the probability ratio between the current and previous policies, $\hat{A}_t$ is the estimated advantage function, and ϵ is a clipping parameter that limits the magnitude of policy updates.

This clipped formulation prevents excessively large policy updates that could otherwise destabilize the learning process, particularly in stochastic and highly constrained environments such as multi-hop LoRa networks. By bounding the probability ratio $r_t(\theta)$ within the interval $[1-\epsilon, 1+\epsilon]$, PPO ensures that newly updated policies do not deviate significantly from the previously deployed policy, thereby preserving stable and incremental policy improvement. This property is especially important in routing optimization, where abrupt changes in forwarding decisions may lead to sudden congestion, increased collision rates, or uneven energy depletion among relay nodes. Fig 3 PPO architecture.

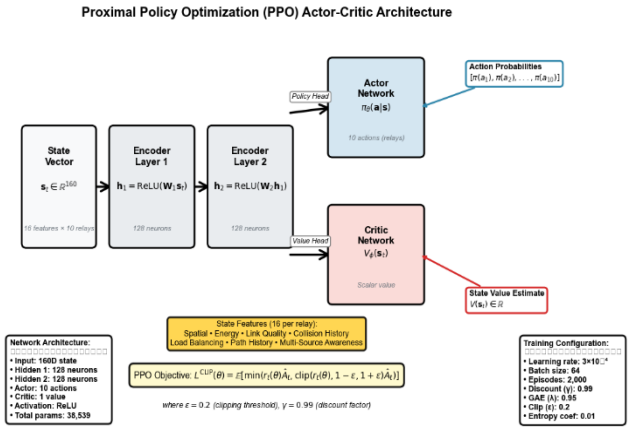


Fig. 3. Proximal Policy Optimization (PPO) architecture.

Algorithm 2 Proximal Policy Optimization (PPO)

INPUT: State s ∈ R^160, Actions A = {a1, ..., a10}, Learning rate α = 3×10^-4
OUTPUT: Optimized routing policy π_θ

NETWORKS:
Actor: π_θ(a|s) = Softmax(f_θ(s))
Critic: V_φ(s) = g_φ(s)

TRAINING (Per Episode):

Action Selection with Temperature Scaling

p(a|s) = (π _ θ (a|s) ⊙ M) / Σ(π _ θ (a|s) ⊙ M)
a_t ~ Categorical(p(a|s)^(1/T))
where M is an action mask, T is the temperature

Discounted Returns and Normalization
G_t = Σ(τ = t to T_end) γ^(τ−t) r_τ
Ĝ_t = (G_t − μ(G)) / (σ(G) + ε_norm)

PPO Update (K = 4 epochs, using collected trajectories)
ratio_t = π_θ(a_t|s_t) / π_θ_old(a_t|s_t)
A_t = Ĝ_t − V_φ(s_t)

L^CLIP = −E[min(ratio_t·A_t , clip(ratio_t, 1−ε, 1+ε)·A_t)]
L^VF = E[(V_φ(s_t) − Ĝ_t)^2]
L^ENT = −E[Σ_a π_θ(a|s_t) log(π_θ(a|s_t))]

L_total = L^CLIP + 0.5·L^VF − 0.01·L^ENT

Update parameters:
θ, φ ← θ, φ − α ∇_{θ,φ} L_total

Gradient Clipping
||∇_{θ,φ}|| ⩽ 0.3

PARAMETERS:
γ = 0.99, ε = 0.2, K = 4,
T = max(0.01, 0.6 − (episode / N_max)·0.59)

C. Training Procedure

Training is conducted over multiple episodes, each consisting of concurrent packet transmissions from multiple source nodes. During training, stochastic action sampling is employed to promote exploration, while temperature scaling is used to avoid premature convergence. Experience tuples (s_t, a_t, r_t, s_{t+1}) are collected across episodes and used to update the policy and value networks in mini batches. To reduce synchronization-induced collisions and encourage path diversity during early learning, guided exploration is applied in initial training phases. As training progresses, the agent gradually transitions to fully autonomous decision-making based on learned policies.

By formulating multi-hop LoRa routing as a reinforcement learning problem and incorporating collision-aware state features and energy-sensitive rewards, the proposed PPO-based framework learns routing policies that mitigate relay hotspots, reduce collision-induced retransmissions, and improve the overall energy–reliability trade-off.

D. Simulation Setup & Configuration

This work is simulation-based rather than hardware-deployed a deliberate methodological choice given that deploying a 27-node multi-hop LoRa mesh (3 sources, 23 relays with 90000mJ battery in a 100 m hexagonal grid, 1 gateway) across a 1.9 km × 1.1 km forested area with a maximum link range of 300 m and a source-to-gateway distance of 1,100 m for controlled, repeatable experiments is logistically and financially impractical within a single study. Simulation allows systematic isolation of individual variables (SF assignment, relay energy, channel occupancy, multi-source contention) under conditions grounded in empirically validated forest propagation parameters. All experiments are implemented in Python 3.11 using PyTorch for the PPO agent, LoRaSim for discrete-event network dynamics, SimPy for real-time scheduling, and NumPy/Pandas for data processing (Table II).

TABLE II. SIMULATION PARAMETERS

LoRa Parameters	**Values**
Gain	2 dBi
CR, c	1
BW	125kHz
Duty Cycle, n	10%
Length of payload, npl	100 bytes
References distance, d0	1m
Path loss index, y	3.8
Shadow fading, x0 `~N	N(0, 8^2) dB
SF	{7, 8, 9, 10, 11, 12}
TP	{8, 11, 14, 17, 20} dBm
Training Parameters	**Values**
State Dimension	160
Action Dimension	10
Learning Rate, α	3×10^{-4}
Total Training Episodes	2000
Discount Factor, γ	0.99
Clip Epsilon, ε	0.2

E. Baseline Routing Methods

To objectively evaluate the effectiveness of the proposed PPO-based routing framework, its performance is compared against three conventional routing strategies: Shortest Path routing, Random routing, and an AODV-like routing method. These baselines are selected to represent commonly used heuristic, stochastic, and protocol-inspired approaches in multi-hop wireless networks.

1) Shortest Path Routing selects the next-hop relay that minimizes the Euclidean distance to the gateway [43]. At each forwarding decision, the relay j^* is chosen such that

$$j^* = \arg\min_{j \in \mathcal{N}_i} d(j, \text{GW}) \quad (22)$$

where $\mathcal{N}_i$ denotes the set of candidates neighboring relays of node i, and $d(j, \text{GW})$ is the Euclidean distance between relay j and the gateway. This strategy minimizes hop distance toward the destination but does not consider relay energy, traffic load, or channel interference. As a result, it often concentrates traffic on a small subset of relays, leading to energy hotspots and increased collision probability.

2) Random Routing selects the next-hop relay uniformly at random from the set of reachable neighboring relays [44]. If N candidate relays are available, the selection probability for each relay j is given by

$$P(j) = \frac{1}{|\mathcal{N}_i|} \quad (23)$$

where $|\mathcal{N}_i|$ is the number of available candidate relays. This baseline does not exploit any network state information and serves as a lower-bound reference for routing performance. While Random routing can occasionally provide path diversity, it generally leads to inefficient routing paths, excessive hop counts, and elevated energy consumption.

3) AODV-like Routing is inspired by the AODV protocol [45] and selects the next-hop relay based on a composite routing metric that accounts for distance to the gateway, residual energy, and relay load. For each candidate relay j, a routing cost function is defined as

$$C(j) = \alpha\ d(j, \mathrm{GW}) + \beta\left(1 - \frac{E_j}{E_{max}}\right) + \gamma\, L_j \qquad (24)$$

where E_j denotes the residual energy of relay j, E_{max} is the initial battery capacity, and L_j represents the relay load, modeled as the ratio of forwarded packets to received packets. The weighting coefficients $\alpha, \beta,$ and γ balance progress toward the gateway, energy awareness, and load distribution, respectively. The relay with the minimum cost $C(j)$ is selected as the next hop. Although the AODV-like method introduces energy and load awareness, it relies on static weighting factors and lacks adaptability to dynamic interference and collision conditions. Consequently, it remains susceptible to suboptimal routing decisions under time-varying traffic and channel states.

V. Simulation Results

This section presents a comprehensive evaluation of the proposed PPO-based multi-hop LoRa routing framework with dynamic ADR and compares its performance against conventional baseline routing strategies, namely Shortest Path, Random Routing, and AODV-like routing. The evaluation focuses on energy efficiency, packet delivery reliability, collision behaviour, and path diversity under realistic propagation and interference conditions.

A. Training Convergences and evaluation

All the figures below illustrate the learning dynamics and performance evolution of the proposed PPO-based routing framework in a multi-hop LoRa network. The results demonstrate how reinforcement learning progressively improves routing decisions by reducing collisions, increasing packet delivery reliability, and reducing energy consumption.

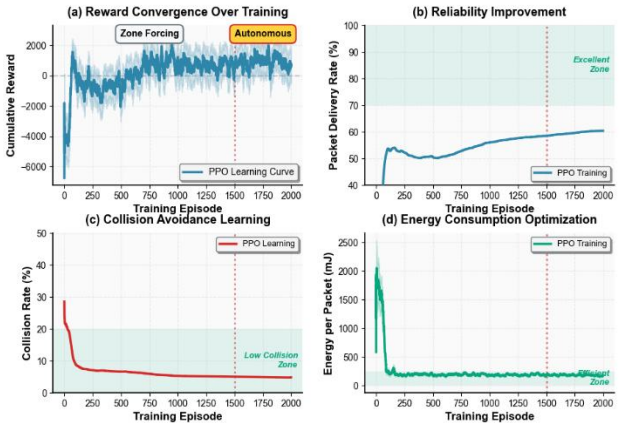


Fig. 4. PPO Learning curve & training optimization.

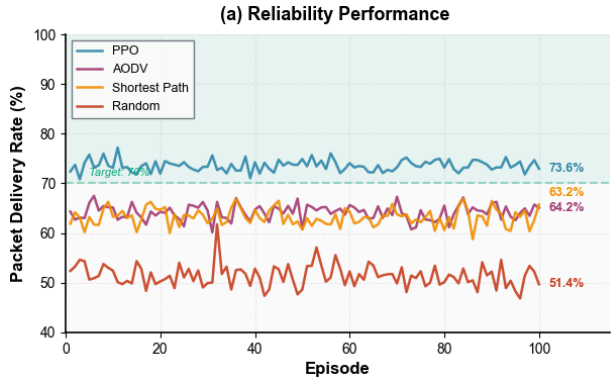


Fig. 5. Packet delivery reliability.

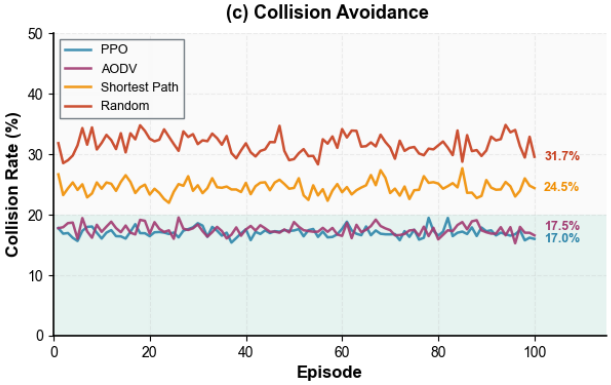


Fig. 6. Collision avoidance.

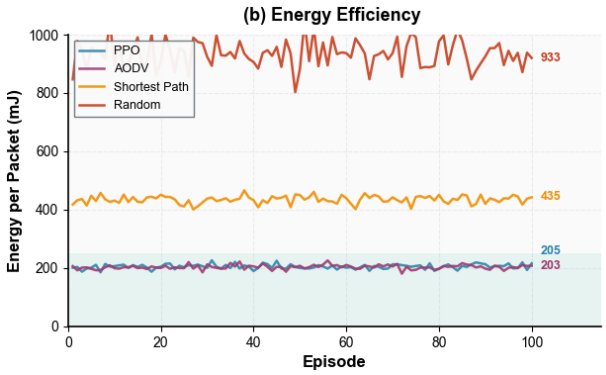


Fig. 7. Energy consumption optimization.

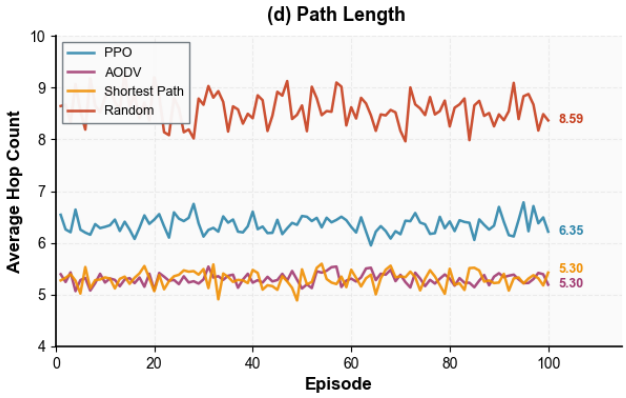


Fig. 8. Average hop count.

The cumulative reward in Fig. 4. represents the expected discounted return $E[\sum_t \gamma^t r_t]$, where r_t penalizes transmission energy E, hop count H, and path overlap O while rewarding successful packet delivery. During early training, the reward is highly negative and exhibits large variance due to exploratory routing decisions that frequently select congested relays, leading to collisions and retransmissions. As training progresses, the PPO policy updates stabilize through clipped probability ratio optimization, resulting in steadily increasing and eventually convergent reward values. The reduction in reward variance indicates policy convergence and reduced exploration.

Fig. 5. shows the evaluation of packet delivery ratio. In LoRa networks, packet decoding succeeds only when the signal-to-interference-plus-noise ratio satisfies $\text{SINR} = \frac{P_r}{\sum I+N} \geq \text{SINR}_{\text{min}}(\text{SF})$ here $\sum I$ represents aggregated co-spreading-factor interference. At early training stages, frequent overlapping transmissions violate this condition, causing packet loss. As the PPO agent learns to avoid relays with high channel occupancy and recent usage, $\sum I$ decrease, allowing more packets to satisfy the SINR constraint. Consequently, the PPO-based approach achieves a higher delivery ratio than Random, AODV-like, and Shortest Path routing.

Average Collision rate is illustrated in Fig. 6. Baseline routing methods, particularly Shortest Path routing, repeatedly select the same relays, increasing the probability of overlapping transmissions and persistent collision hotspots. In contrast, the PPO agent minimizes path overlap O by incorporating relay usage history and collision statistics into the state representation. This leads to a rapid reduction in collision rate during training and a significantly lower steady-state collision level compared with all baseline methods.

Fig. 7. presents the average energy consumption per delivered packet. The transmission energy is modeled as $E_{tx} = I_{tx} \times V \times T_{air}$, where T_{air} increases exponentially with the spreading factor. During early training, frequent retransmissions and ADR-triggered increases in spreading factors result in excessive energy consumption. As collision probability decreases under PPO routing, retransmissions are reduced and ADR operates at lower spreading factors and transmit power levels, thereby reducing both T_{air} and transmission current. The combined effect yields the lowest energy consumption per delivered packet among all evaluated methods.

Average Hop count in Fig. 8 shows PPO routes are longer than Shortest Path and AODV, but far shorter than Random routing. PPO learns a controlled trade-off: it accepts a modest hop-count increase (about 6.35 hops on average) to bypass relays that are frequently selected by other sources, which are more likely to be congested. In a contention-limited LoRa mesh, avoiding these hotspot relays reduces co-SF airtime overlap, lowering collision probability and the need for energy-costly retransmissions. Illustrate in Fig. 10, PPO still preserves forward progress toward the gateway, so the additional hops reflect purposeful path diversification rather than inefficient wandering.

Fig. 4 also shows the PPO training convergence across 2,000 episodes in two phases. The agent initially produces highly negative cumulative rewards (−6,000) due to random exploration, then improves rapidly between episodes 50–500 as the policy learns interference avoidance and load balancing. Stable performance is reached at approximately episode 500 and sustained through the Autonomous phase (episodes 1,500–2,000) without performance collapse, confirming the policy generalizes beyond the structured curriculum. Training is performed entirely offline; deployed relay nodes execute only the inference step a single forward pass through a compact neural network making the approach fully compatible with resource-constrained LoRa hardware. Overall, these results confirm that energy efficiency and reliability in multi-hop LoRa networks are dominated by collision-induced retransmissions rather than hop count alone, and that the PPO-based framework outperforms conventional heuristic routing across all performance metrics by learning interference-aware, load-balanced policies that cooperate implicitly with ADR.

B. Path Diversity and Spatial Traffic Separation

Figures below illustrate the routing behaviour of multiple source nodes before and after PPO learning, highlighting the impact of learned routing policies on spatial traffic separation and collision reduction. Prior to PPO learning, routing paths from different sources converge toward the same set of relays that are geographically closest to the gateway. This convergence creates persistent collision hotspots, as multiple packets are forwarded simultaneously through the same relays. In LoRa networks, when multiple transmissions using the same spreading factor overlap in time, the aggregated interference power ∑I increases, and packet decoding fails if the signal-to-interference-plus-noise ratio, defined as $SINR = P_r/(sumI + N)$, falls below the spreading-factor-dependent threshold SINR_{min} a result, the convergent routing behavior observed before learning leads to a high collision rate of approximately 34.7%. As shown in Fig. 9.

After PPO learning as shown in Fig. 10, the routing paths exhibit clear spatial separation, with each source predominantly selecting relays in distinct regions of the network. This behaviour emerges from the PPO reward formulation, which penalizes path overlap O and collision-induced retransmissions while rewarding successful packet delivery. By observing relay usage history, channel occupancy, and recent transmission activity in the state representation, the PPO agent learns to minimize O, defined as the number of concurrent routing decisions that reuse the same relay within a short time window. Reducing O directly lowers the probability of overlapping transmissions, thereby decreasing the aggregated interference ∑I and increasing the likelihood that the SINR condition for successful decoding is satisfied.

The spatial separation of routing paths also improves energy efficiency and relay fairness. Since the energy consumed per transmission is modelled $E_{tx} = I_{tx} \times V \times T_{air}$, and retransmissions incur the same energy cost as successful transmissions, reducing collision probability significantly lowers total energy consumption. Furthermore, by distributing traffic across multiple relays rather than repeatedly using the same nodes, PPO prevents energy hotspots and slows battery depletion at individual relays. This balanced relay utilization contributes to improved network sustainability and longer operational lifetime.

Overall, the results demonstrate that path diversity is not achieved through random exploration but through learned, interference-aware routing decisions. The PPO-based framework effectively transforms routing behaviour from convergence-driven interference patterns into spatially distributed paths that reduce collisions, improve reliability, and enhance energy efficiency in multi-hop LoRa networks.

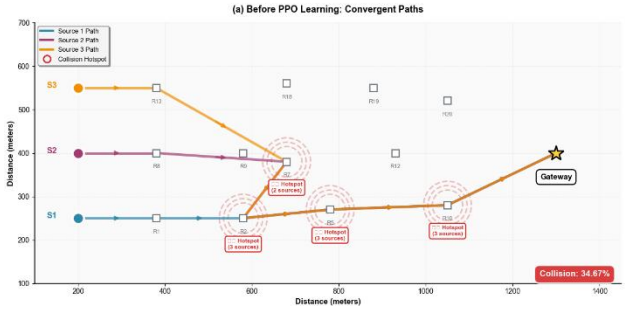


Fig. 9. Before spatial traffic separation.

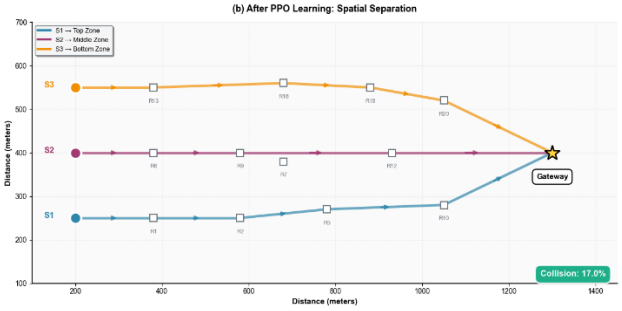


Fig. 10. After spatial traffic separation.

VI. Conclusion

This paper proposes an energy-efficient and reliable routing framework for multi-hop LoRa networks using PPO with dynamic ADR. A discrete-event simulation models realistic LoRa physical-layer effects, co-spreading-factor interference, battery-limited relays, and multi-source contention. Unlike conventional hop-count or heuristic routing, the method formulates routing as a sequential decision problem and learns interference-aware, load-balanced policies via reinforcement learning. Results show significant improvements in packet delivery reliability, collision reduction, and energy consumption per delivered packet compared with Shortest Path, Random routing, and an AODV-like baseline. The learned policy sacrifices minimal hop count to reduce interference and retransmissions, promotes spatial path diversity, mitigates relay energy hotspots, slows battery depletion, and enhances network sustainability.

Overall, energy efficiency is driven primarily by collision-induced retransmissions rather than hop count alone. By accounting for interference, relay load, and transmission history, the PPO-based framework demonstrates the promise of reinforcement learning for interference-limited IoT mesh routing, with physical hardware validation remaining an important direction for future work.

A key limitation of this work is the sim-to-real gap. The propagation model captures the dominant effects of forest path loss, vegetation attenuation, and log-normal shadow fading based on empirically validated parameters [5], [6], [32], [46] however, real deployments additionally involve hardware specific noise figures, adjacent-channel interference, oscillator frequency drift, and the discrete timing constraints of the LoRa duty-cycle and Channel Activity Detection (CAD) mechanism. To reduce sensitivity to these differences, the PPO policy operates on normalised relative state features residual energy ratios, channel-occupancy indicators, and hop-count differences rather than raw absolute signal values, which improves robustness to hardware-to-hardware calibration offsets. Nevertheless, transfer validation on physical LoRa nodes in a controlled forest testbed remains a necessary step before operational deployment.

Acknowledgment

The authors would like to acknowledge Universiti Malaysia Sabah (UMS), Ministry of Higher Education Malaysia (KPT), and Ministry of Science, Technology and Innovation (MOSTI) for supporting this research under Fundamental Research Grant Scheme - Early Career Researcher (FRGS-EC), grant no. FRGS-EC/1/2024/TK07/UMS/02/3 (university code: FRGC056-2024) and Strategic Research Fund (SRF) – Programme Space based Technology, grant no. LPK2414.

References

[1] K. Shafique, B. A. Khawaja, F. Sabir, S. Qazi, and M. Mustaqim, "Internet of things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT Scenarios," 2020, Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ACCESS.2020.2970118.

[2] 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.2017.12.005.

[3] F. Yao, Y. Ding, S. Hong, and S.-H. Yang, "A Survey on Evolved LoRa-Based Communication Technologies for Emerging Internet of Things Applications," International Journal of Network Dynamics and Intelligence, pp. 4–19, Dec. 2022, doi: 10.53941/ijndi0101002.

[4] A. I. Griva et al., "LoRa-Based IoT Network Assessment in Rural and Urban Scenarios," Sensors, vol. 23, no. 3, Feb. 2023, doi: 10.3390/s23031695.

[5] G. Callebaut and L. Van Der Perre, "Characterization of LoRa Point-to-Point Path Loss: Measurement Campaigns and Modeling Considering Censored Data," IEEE Internet Things J., vol. 7, no. 3, pp. 1910–1918, Mar. 2020, doi: 10.1109/JIOT.2019.2953804.

[6] R. El Chall, S. Lahoud, and M. El Helou, "LoRaWAN network: Radio propagation models and performance evaluation in various environments in Lebanon," IEEE Internet Things J., vol. 6, no. 2, pp. 2366–2378, Apr. 2019, doi: 10.1109/JIOT.2019.2906838.

[7] A. Lahoua, M. Younes, and L. Touseau, "Improving Energy-Efficiency In Lora Networks Thanks to a Holistic Transmission Model," pp. 469–476, 2024, doi: 10.1109/WiMob61911.2024.10770422ï.

[8] D. Zorbas, C. Caillouet, K. A. Hassan, and D. Pesch, "Optimal data collection time in lora networks— a time-slotted approach," Sensors (Switzerland), vol. 21, no. 4, pp. 1–22, Feb. 2021, doi: 10.3390/s21041193.

[9] D. L. Mai and M. K. Kim, "Multi-hop LOra network protocol with minimized latency," Energies (Basel)., vol. 16, no. 3, Mar. 2020, doi: 10.3390/en13061368.

[10] R. Airiyoshi, M. Hasegawa, T. Ohtsuki, and A. Li, "Energy Efficient Transmission Parameters Selection Method Using Reinforcement Learning in Distributed LoRa Networks," Jan. 2025, [Online]. Available: http://arxiv.org/abs/2410.11270

[11] A. Lahoua, M. Younes, and L. Touseau, "Improving Energy-Efficiency In Lora Networks Thanks to a Holistic Transmission Model," pp. 469–476, 2024, doi: 10.1109/WiMob61911.2024.10770422ï.

[12] M. Alkhayyal and A. M. Mostafa, "Enhancing LoRaWAN Sensor Networks: A Deep Learning Approach for Performance Optimizing and Energy Efficiency," Computers, Materials and Continua, vol. 83, no. 1, pp. 1079–1100, 2025, doi: 10.32604/cmc.2025.061836.

[13] X. Chen, Y. Mao, Y. Xu, W. Yang, C. Chen, and B. Lei, "Energy-efficient multi-hop LoRa broadcasting with reinforcement learning for IoT networks," Ad Hoc Networks, vol. 169, Mar. 2025, doi: 10.1016/j.adhoc.2024.103729.

[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," Aug. 2017, [Online]. Available: http://arxiv.org/abs/1707.06347

[15] R. Serati, B. Teymuri, N. A. Anagnostopoulos, and M. Rasti, "ADR-Lite: A Low-Complexity Adaptive Data Rate Scheme for the LoRa Network," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.14583

[16] N. Izzeldin, M. Osman, and E. B. Abbas, "Performance Evaluation of LoRa and Sigfox LPWAN Technologies for IoT," Academic Journal of Research and Scientific Publishing |, vol. 4, pp. 5–6, [Online]. Available: www.ajrsp.comwww.ajrsp.com

[17] M. Lauridsen, H. Nguyen, B. Vejlgaard, I. Z. Kovacs, P. Mogensen, and M. Sorensen, "Coverage Comparison of GPRS, NB-IoT, LoRa, and SigFox in a 7800 km Area," in IEEE Vehicular Technology Conference, Institute of Electrical and Electronics Engineers Inc., Nov. 2017. doi: 10.1109/VTCSpring.2017.8108182.

[18] A. Loubany, S. Lahoud, A. E. Samhat, and M. El Helou, "Improving Energy Efficiency in LoRaWAN Networks with Multiple Gateways," Sensors, vol. 23, no. 11, Jun. 2023, doi: 10.3390/s23115315.

[19] M. Bor, U. Roedig, T. Voigt, and J. M. Alonso, "Do LoRa low-power wide-area networks scale?," in MSWiM 2016 - Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, Association for Computing Machinery, Inc, Nov. 2016, pp. 59–67. doi: 10.1145/2988287.2989163.

[20] S. Feng, J. Chen, and Z. Zhao, "Cost Effective Routing in Large-scale Multi-hop LoRa Networks," in INFOCOM WKSHPS 2022 - IEEE Conference on Computer Communications Workshops, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/INFOCOMWKSHPS54753.2022.9798302.

[21] F. Alghamdi and F. Bajaber, "Adaptive Real-Time Channel Estimation and Parameter Adjustment for LoRa Networks in Dynamic IoT Environments," Sensors, vol. 25, no. 7, Apr. 2025, doi: 10.3390/s25072121.

[22] L. Networks, st Yi Jiang, nd Minghua Wang, and rd Xingbin Wang, "A Efficient Adaptive Data Rate Algorithm in," 2023.

[23] M. Coutinho, J. A. Afonso, and S. F. Lopes, "An Efficient Adaptive Data-Link-Layer Architecture for LoRa Networks," Future Internet, vol. 15, no. 8, Aug. 2023, doi: 10.3390/fi15080273.

[24] M. Misbahuddin, M. S. Iqbal, and L. A. S. I. Akbar, "A Multi-Hop Routing Solution for Low Latency and Energy Efficiency in Large-Scale LoRa IoT," International Journal of Computers, Communications and Control, vol. 20, no. 3, 2025, doi: 10.15837/ijccc.2025.3.6501.

[25] R. Pueyo Centelles, R. Meseguer, F. Freitag, R. Baig Viñas, and L. Navarro, "Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000. A minimalistic distance-vector routing protocol for LoRa mesh networks", doi: 10.1109/ACCESS.2017.DOI.

[26] L. Acosta-Garcia, J. Aznar-Poveda, A. J. Garcia-Sanchez, J. Garcia-Haro, and T. Fahringer, "Dynamic transmission policy for enhancing LoRa network performance: A deep reinforcement learning approach," Internet of Things (Netherlands), vol. 24, Dec. 2023, doi: 10.1016/j.iot.2023.100974.

[27] J. Haxhibeqiri, F. Van den Abeele, I. Moerman, and J. Hoebeke, "LoRa scalability: A simulation model based on interference measurements," Sensors (Switzerland), vol. 17, no. 6, Jun. 2017, doi: 10.3390/s17061193.

[28] I. Urabe, A. Li, M. Fujisawa, S. J. Kim, and M. Hasegawa, "Combinatorial MAB-Based Joint Channel and Spreading Factor Selection for LoRa Devices," Sensors, vol. 23, no. 15, Aug. 2023, doi: 10.3390/s23156687.

[29] F. Alghamdi and F. Bajaber, "Adaptive Real-Time Channel Estimation and Parameter Adjustment for LoRa Networks in Dynamic IoT Environments," Sensors, vol. 25, no. 7, Apr. 2025, doi: 10.3390/s25072121.

[30] A. E. Ferreira et al., "A study of the LoRa signal propagation in forest, urban, and suburban environments", doi: 10.1007/s12243.

[31] Y. Wu, G. Guo, G. Tian, and W. Liu, "A Model with Leaf Area Index and Trunk Diameter for LoRaWAN Radio Propagation in Eastern China Mixed Forest," J. Sens., vol. 2020, 2020, doi: 10.1155/2020/2687148.

[32] M. R. Ansah, R. A. Sowah, J. Melià-Seguí, F. A. Katsriku, X. Vilajosana, and W. O. Banahene, "Characterising foliage influence on LoRaWAN pathloss in a tropical vegetative environment," Oct. 01, 2020, Institution of Engineering and Technology. doi: 10.1049/iet-wss.2019.0201.

[33] S. Rougerie, J. Israel, K. Tomoshige, S. Rougerie, J. Israel, and T. Kan, "Validation of ITU-R P.833-9 tree attenuation model for Land Mobile Satellite propagation channel at Ku/Ka band." [Online]. Available: https://hal.science/hal-03231346v1

[34] Z. Xu, S. Tong, P. Xie, and J. Wang, "From Demodulation to Decoding: Toward Complete LoRa PHY Understanding and Implementation," ACM Trans. Sens. Netw., vol. 18, no. 4, Jan. 2023, doi: 10.1145/3546869.

[35] H. Yang, H. Ji, Z. Huang, and X. Wu, "A GNN-Based Learning Approach for Energy Optimization in Relay-Assisted IoT Networks," in IEEE Wireless Communications and Networking Conference, WCNC, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/WCNC61545.2025.10978775.

[36] N. Abdoun, S. Abboud, H. Altaleb, Z. Rajnai, and B. Donát, "Collision Detection in LoRaWAN Using Machine Learning," pp. 1–6, 2024, doi: 10.1109/sisy62279.2024.10737620ï.

[37] Y. A. Al-Gumaei, N. Aslam, M. Aljaidi, A. Al-Saman, A. Alsarhan, and A. Y. Ashyap, "A Novel Approach to Improve the Adaptive-Data-Rate Scheme for IoT LoRaWAN," Electronics (Switzerland), vol. 11, no. 21, Nov. 2022, doi: 10.3390/electronics11213521.

[38] F. Bizzarri, C. Mocenni, and S. Tiezzi, "A Markov Decision Process with Awareness and Present Bias in Decision-Making," Mathematics, vol. 11, no. 11, Jun. 2023, doi: 10.3390/math11112588.

[39] M. Löppenberg, S. Yuwono, M. R. Diprasetya, and A. Schwung, "Dynamic robot routing optimization: State–space decomposition for operations research-informed reinforcement learning," Robot. Comput. Integr. Manuf., vol. 90, Dec. 2024, doi: 10.1016/j.rcim.2024.102812.

[40] Z. Sun, H. Tian, H. Wang, S. Yan, T. Han, and H. Rong, "Utilising a two-dimensional action space as the basis for the PPO algorithm, job shops can jointly schedule AGVs and machinery," Sep. 24, 2025. doi: 10.21203/rs.3.rs-7413549/v1.

[41] M. Zheng, J. Zhang, C. Zhan, X. Ren, and S. Lü, "Proximal policy optimization with reward-based prioritization," Expert Syst. Appl., vol. 283, Jul. 2025, doi: 10.1016/j.eswa.2025.127659.

[42] K. Sun, J. Yang, J. Li, B. Yang, and S. Ding, "Proximal Policy Optimization-Based Hierarchical Decision-Making Mechanism for Resource Allocation Optimization in UAV Networks," Electronics (Switzerland), vol. 14, no. 4, Feb. 2025, doi: 10.3390/electronics14040747.

[43] H. M. Hadi and I. M. Ibrahim, "A Comprehensive Review of Shortest Path Algorithms for Network Routing," Asian Journal of Research in Computer Science, vol. 18, no. 3, pp. 152–175, Feb. 2025, doi: 10.9734/ajrcos/2025/v18i3584.

[44] D. C. Dhanapala, A. P. Jayasumana, and Q. Han, "On random routing in wireless sensor grids: A mathematical model for rendezvous probability and performance optimization," J. Parallel Distrib. Comput., vol. 71, no. 3, pp. 369–380, Mar. 2011, doi: 10.1016/j.jpdc.2010.10.016.

[45] S. Liu, Y. Yang, and W. Wang, "Research of AODV Routing Protocol for Ad Hoc Networks1," AASRI Procedia, vol. 5, pp. 21–31, 2013, doi: 10.1016/j.aasri.2013.10.054.

[46] A. E. Ferreira et al., "A study of the LoRa signal propagation in forest, urban, and suburban environments", doi: 10.1007/s12243.

Comparative Analysis of Fixed vs Machine Learning Dynamic Pricing Models: A Computational Performance Study

Emmanuel Ofotsu Kwesi Bannor[1]*, S. Sarah Maidin[2],
Vinayakumar Ravi[3], Nguyen Thi Thu Thuy[4], Nghiem Thi-Lich[5]
Department of Data Science and Information Technology, INTI International University, Nilai, Malaysia[1, 2]
Centre for Data Science and Sustainable Technologies-Faculty of Data Science and Information Technology, INTI, International University, 71800 Nilai, Negeri Sembilan, Malaysia[2]
Department of IT and Methodology, Wekerle Sandor Uzleti Foiskola, Budapest, Hungary[2, 3]
Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia[4]
Faculty of Economic Information System and E-Commerce, Thuongmai University, Hanoi, Vietnam[5]

***Abstract*—The rise of e-commerce and digital offerings has generated a need for ultra-adaptable pricing policies seeking to maximize revenue while optimizing competitive advantage. Traditional fixed pricing schemes are inherently flawed due to a lack of responsiveness to instantaneous fluctuations in the marketplace, inventory levels, as well as demand inelasticity. This study conducts a detailed computational performance study comparing fixed pricing, standard heuristic dynamic pricing (HDP), advanced Machine Learning (ML)-oriented dynamic pricing schemes, with a special focus on a Bi-LSTM network as well as a hybrid scheme based on Wavelet Decomposition (WD). Through simulated high-frequency transactions as well as marketplace data, model evaluation relies on three critical performance metrics: Total Revenue Generated, Pricing Accuracy (measured through Mean Absolute Percentage Error, MAPE), as well as Computational Latency (vital for real-time utilization). The results indicate that while HDP shows marginal improvements over fixed pricing, ML-based schemes, particularly a hybrid WD-Bi-LSTM model, exhibit substantial revenue maximization (up to 18.5% improvement) as well as forecasting accuracy (MAPE up to 2.1%), though at a slight increase in computational latency remains acceptable for real-time deployment for near real-time deployment. This study provides a quantitative foundation for organizations embracing AI-supportive pricing initiatives with emphasis on trade-offs among model sophistication, predictive potency, as well as functionality performance.**

***Keywords*—Computational performance; deep learning; dynamic pricing; Machine Learning (ML); process innovation**

I. Introduction

In the contemporary digital economy, pricing remains a primary lever for profit maximization and market positioning. The conventional approach of fixed pricing, determined through cost-plus or value-based analysis, is increasingly becoming obsolete as it fails to capture the intricate, rapidly changing dynamics of demand, competitor actions, and external economic factors [1], [8]. While fixed pricing remains widely used in retail and subscription-based markets, industry reports indicate that over 60–70% of large-scale e-commerce platforms now employ some form of dynamic pricing to respond to demand volatility and competitive pressures. However, fixed pricing continues to be effective in low-volatility environments and regulated industries. Therefore, rather than being obsolete, fixed pricing serves as a baseline benchmark against which adaptive pricing strategies can be evaluated quantitatively.

The shift towards dynamic pricing where prices are adjusted in real-time or near real-time is not merely an operational optimization but a strategic imperative [2]. The computational challenge of dynamic pricing lies in its reliance on accurate short-term demand and price forecasting, which is highly influenced by diverse, non-linear variables, including sentiment derived from product descriptions and financial behaviors [4], [5]. Early dynamic pricing strategies relied on simple rule-based heuristics, which, while flexible, lacked the predictive power needed to optimize complex, high- dimensional data [2]. The advent of Artificial Intelligence (AI) and Machine Learning (ML) has fundamentally transformed this landscape by offering sophisticated algorithms capable of processing voluminous data streams and inferring optimal price points that reflect true market equilibrium [8]. This paper addresses a critical gap in the existing literature by providing a detailed comparative analysis of the computational efficiency and operational performance of three distinct pricing paradigms:

- Fixed Pricing Model (FP): A baseline model using a static price.
- Heuristic Dynamic Pricing Model (HDP): A rule-based model utilizing simple inventory and time-of- day variables.
- Machine Learning Dynamic Pricing Model (ML- DP): Advanced models, including a pure Bi-LSTM network and a hybrid decomposition-based network, designed for high-accuracy forecasting.

The objective is to quantify the performance differential in terms of revenue, predictive error, and, crucially, the computational latency required for real-time deployment. This comparison is essential for organizations making informed decisions about technology adoption for enhancing competitiveness in e-commerce and related sectors [1].

*Corresponding author.

II. Related Work

The foundation of dynamic pricing is rooted in operations research and control theory, evolving rapidly with advancements in computational capacity. The current state-of-the-art heavily favors ML, particularly Reinforcement Learning (RL) and Deep Learning (DL), for managing complex resource allocation problems where pricing is a critical component [3]. The ability of ML to leverage multi- modal data integration for node classification and pattern recognition underscores its suitability for dynamic market modeling [7].

Traditional predictive models, such as those used for concrete strength prediction or energy consumption forecasting, have demonstrated the superior performance of ML techniques over conventional statistical methods in handling non-linear data [6], [11]. Specifically in the domain of energy price and consumption, models like the two-layer decomposition technique combined with Bi-LSTM (Bidirectional Long Short-Term Memory) networks have shown exceptional short-term forecasting accuracy, which is directly analogous to the challenges of demand forecasting in dynamic pricing [9], [10]. Likewise, cluster-based ensemble methods have been used effectively in analyzing time-series datasets, leading to an understanding of pattern identification, which helps in optimal pricing [11], [12].

Additionally, studies have looked at the use of hybrid models incorporating signal processing (Wavelet Decomposition) and DL, which improves predictions for highly volatile markets, including electricity prices on a day-ahead basis [14]. The hybrid model addresses problems related to noise and non-stationarity prevalent in market datasets [16]. While extensive literature exists on the theoretical optimization benefits of dynamic pricing [16], a focused comparative study detailing the computational overhead against the revenue gain across fixed, heuristic, and sophisticated hybrid ML models, particularly within a unified simulation framework, remains underdeveloped. This paper directly addresses this gap by quantifying the necessary computational resources against the realized economic benefit.

Dynamic pricing research has seen recent developments where reinforcement learning techniques have been applied to optimize pricing strategies. With reinforcement learning algorithms, prices can be determined as part of a sequential decision process, and they work through trial-and-error processes to maximize the total rewards accrued. Studies have demonstrated that Deep Q-Networks (DQN) and policy gradient methods outperform traditional supervised learning approaches in highly dynamic environments with delayed reward structures. However, these models often require substantial training data and computational resources, motivating the need for hybrid forecasting-based approaches as explored in this study [12].

A brief review of the cited literature reveals a pronounced trend, with 77% of the included works published after 2020, confirming the accelerating research interest and application of ML/AI in price optimization, demand management, and energy consumption forecasting [17], [18], [19], [20], [21], [22]. This bibliometric observation validates the timely focus of this comparative study.

III. Methodology

This study adopts a controlled computational experimental framework to compare the operational and predictive performance of fixed pricing, heuristic dynamic pricing, and machine learning-based pricing systems under identical market conditions. All models were evaluated using the same synthetic high-frequency transaction dataset to ensure fairness, consistency, and reproducibility.

A. Research Design

The study employs a controlled, comparative experimental design. Three primary models—Fixed Pricing (FP), Heuristic Dynamic Pricing (HDP), and Machine Learning Dynamic Pricing (ML- DP)—are implemented and run over the same simulated market periods. The performance of each model is evaluated against the key metrics defined in Section IV-A. Table IV illustrates the research design employment process. A six-month transactional simulation environment was developed to emulate realistic e-commerce market behavior. The pricing engine updated prices every 15 minutes, resulting in 96 pricing intervals per day and 17,280 total observations across the study period. At each interval, the following steps were executed:

- Capture current market state variables.
- Generate demand estimate.
- Compute price using assigned pricing model.
- Simulate sales quantity based on elasticity response.
- Update inventory and cumulative revenue.
- Store performance metrics.

The same transactional stream was used across all pricing models, while ten independent runs with different random seeds were conducted to improve reliability. The overall experimental setup follows a structured research design to ensure fair evaluation across all models, from data generation through to performance assessment and statistical comparison. The overall research design and corresponding responsibilities of each model are summarized in Table I.

TABLE I. Research Design Employed

Step	Description	Responsible Model(s)
1	Data Simulation	All
2	Model Training (If applicable)	ML- DP
3	Iterative Pricing/Sale Cycle	All
4	Metric Calculation (Revenue, MAPE, Latency)	All
5	Statistical Comparison	All

The research process can be seen from Table I and is structured into five sequential steps, ranging from data simulation to the performance evaluation of all the models used. Whereas all models take part in almost every step, training is unique to the ML-DP only model due to its nature of being trained through machine learning techniques.

B. Comparative Study Design (Flow Chart)

The experimental procedure for conducting the comparison study is presented in Fig. 1. The figure shows the process flow starting from data generation to training, pricing, and evaluation phases.

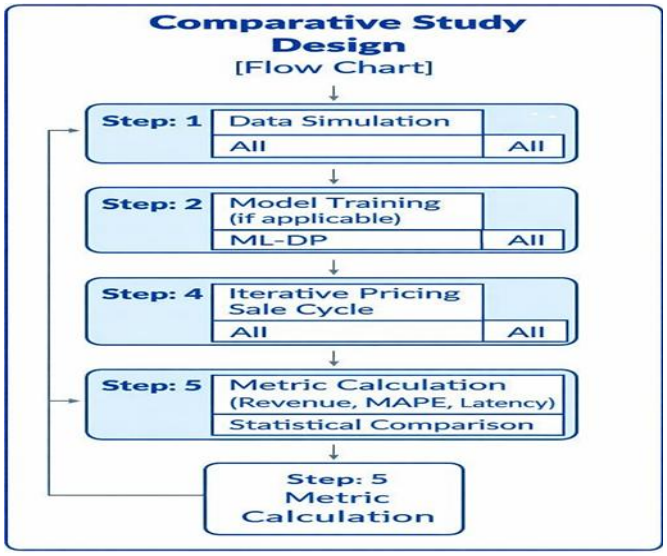


Fig. 1. Comparative study design.

As shown in Fig. 1, the study follows a structured pipeline beginning with data simulation applied across all models, followed by selective training for the ML-DP approach. The iterative pricing-sale cycle is then executed uniformly, and performance is assessed using key metrics such as revenue, Mean Absolute Percentage Error (MAPE), and computational latency, enabling statistical comparison across methods.

C. Data Collection

Given the proprietary nature of real-world dynamic pricing data, a simulated dataset is generated, representative of an e-commerce platform that experiences daily and weekly demand fluctuations.

- *Variables:* Time of day, day of week, inventory level, competitor price, simulated demand elasticity, and historical sales volume.
- *Time Scale:* Data points are generated at 15-minute intervals over a six-month period, mirroring the high-frequency data used in complex energy management and demand response systems [13], [15].

D. Inclusion and Exclusion

- Inclusion Criteria (Models): Models must be fully autonomous (no manual price intervention).
- Exclusion Criteria (Data): Data points corresponding to major public holidays or known exogenous "black swan" events are excluded to maintain a focus on routine operational performance and to enhance the reliability of the forecasting models [17].

E. Machine Learning-Based Dynamic Pricing Models

To enhance forecasting performance in dynamic pricing, two machine learning–based models are employed for comparative evaluation as follows:

- ML Model 1 (Bi-LSTM): A standard deep learning model suitable for sequential data forecasting [9].
- ML Model 2 (WD-Bi-LSTM Hybrid): Incorporates Wavelet Decomposition on the input features to denoise the time series prior to feeding it into the Bi-LSTM network, enhancing prediction for building energy consumption [21].

The proposed WD-Bi-LSTM model applies Daubechies-4 (db4) wavelet decomposition with 3 levels to preprocess transactional demand signals. Approximation and detail coefficients were reconstructed and normalized before input to a two-layer Bi-LSTM network comprising 128 and 64 hidden units, followed by dense layers of 32 and 1 neuron respectively. The dropout rate was fixed at 0.2. The model was trained using Adam optimizer having learning rate = 0.001, batch size = 64, epochs = 100, and patience = 10. The generated demand is log-normally distributed because this probability distribution is commonly applied in demand models due to its property of skewness and heavy tails, similar to those that exist in real-life cases of demand behavior. Calibration was conducted based on e-commerce public benchmarks to provide realistic variance and seasonality of the generated demand.

F. Pricing Policy Function

The pricing decision is derived from predicted demand using a revenue maximization policy:

$$p_t^* = \arg\max_p \; p \cdot \widehat{D}(p, x_t)$$

where, $\widehat{D}$ is the predicted demand generated by the forecasting model. For the heuristic model, price adjustments follow predefined rules based on inventory thresholds, while the ML-based models use gradient-based optimization to determine the optimal price at each time step.

G. Problem Formulation

The dynamic pricing task is formulated as a constrained optimization problem aimed at maximizing expected revenue over a finite time horizon:

$$\max_{p_t} \; R = \sum_{t=1}^{T} p_t \cdot D(p_t, x_t)$$

Subject to:

$$p_{min} \le p_t \le p_{max}, I_t \ge 0$$

where, p_t denotes the price at time t, $D(\cdot)$ represents demand as a function of price and contextual variables x_t, and I_t represents inventory levels. This formulation captures the trade-off between revenue maximization, demand responsiveness, and operational constraints.

H. Hyperparameter Tuning

To ensure optimal model performance and avoid overfitting, a systematic hyperparameter tuning process was conducted using grid search on the validation dataset. Multiple combinations of key training parameters were evaluated to identify configurations that balance predictive accuracy and computational efficiency. The parameters explored include network architecture, regularization, and training dynamics, as summarized in Table II.

TABLE II. HYPERPARAMETER SEARCH SPACE FOR MODEL OPTIMIZATION

Parameter	Values Tested
LSTM Units	64, 128, 256
Layers	1, 2, 3
Dropout	0.1, 0.2, 0.3
Batch Size	32, 64, 128
Learning Rate	0.01, 0.001, 0.0005
Epochs	50, 100, 150

As can be seen from Table II, various values of hyperparameters were used in order to obtain different levels of complexity and training behavior of the model. Low learning rate and intermediate batch size had a positive effect on the stability of convergence, while the use of deep neural networks had a positive impact on the capacity of the model, but at the same time required more computing resources.

I. Final Selected Parameters

After conducting the process of hyperparameter optimization, the best set of parameters was chosen considering both validation accuracy and computation time. The table below represents the model parameters. For this particular arrangement, the number of neurons is taken to be 128 and 64 in two layers, respectively, along with the dropout parameter set at 0.2, the batch size at 64, and the learning rate at 0.001. Such a hyperparameter arrangement strikes the right balance between prediction accuracy and computational efficiency for dynamic pricing applications.

TABLE III. FINAL OPTIMAL HYPERPARAMETER CONFIGURATION

Parameter	Final Value
Layers	2
Units	128, 64
Dropout	0.20
Batch Size	64
Learning Rate	0.001
Epochs	100

From Table III above, it is evident that the chosen configuration uses a two-layer model, whereby each layer reduces the number of units. This makes it possible to extract features from data while controlling for model complexity. The optimal choice of the dropout factor prevents overfitting, whereas the learning rate and batch size facilitate stable convergence.

J. Evaluation Metrics

Model performance was quantified using three core evaluation metrics:

- *Total Revenue Generated ($):* This represents the primary economic performance indicator and was calculated as the cumulative sum of all transactions over the simulation period:

$$\text{Total Revenue} = \sum_{t=1}^{n} (Price_t \times QuantitySold_t)$$

- *Pricing Accuracy (MAPE):* Mean Absolute Percentage Error (MAPE) was used to evaluate the forecasting precision of the underlying demand prediction models employed in the HDP and ML- DP approaches. Lower MAPE values indicate higher predictive accuracy:

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{\text{ActualDemand}(t) - \text{PredictedDemand}(t)}{\text{ActualDemand}(t)} \right| \times 100$$

- *Computational Latency (µs):* This metric measures the average time required for each model to process the current market state and generate an optimal pricing decision. It is a critical indicator for real-time deployment, where rapid response times are essential for operational efficiency and scalability.

K. Data Analysis

The data analysis involves:

- Descriptive Statistics: Summarizing the total revenue and mean computational latency for each model.
- Inferential Statistics: Applying a one-way ANOVA test to determine if the differences in mean revenue and latency across the three model categories (FP, HDP, ML-DP) are statistically significant.

L. Validation and Reliability

To ensure the validity of the ML model results, a ***K***- fold cross-validation (k=5) is performed on the training data. The reliability is further assessed by repeating the full simulation cycle 10 times with different random seeds for demand generation and reporting the average and standard deviation of all metrics [18]. The robust performance of models, such as those for chiller energy consumption, provides a benchmark for expected prediction reliability [19].

M. Ethical Considerations

While the study is based on simulated data, the ethical implications of ML-driven dynamic pricing are acknowledged. Specifically, the Machine learning-based dynamic pricing models are audited against potential feedback loops that could lead to price discrimination based on non-market factors, ensuring compliance with fairness principles [22].

IV. RESULTS AND DISCUSSION

This section presents the quantified performance metrics of the Fixed Pricing, Heuristic Dynamic Pricing, and Machine Learning–based dynamic pricing models across the simulated environment.

A. Economic and Predictive Performance

Table IV summarizes the key economic and predictive metrics over the full six-month simulation period.

TABLE IV. ECONOMIC AND PREDICTIVE PERFORMANCE COMPARISON (6-MONTH SIMULATION)

Model Type	Total Revenue Generated ($)	ΔRevenue vs. Fixed Pricing	MAPE (%)	Std Dev (Revenue, $)
Fixed Pricing (FP)	10,000,000	Baseline	N/A	450,000
Heuristic Dynamic Pricing (HDP)	11,250,000	+12.50%	14.5%	480,000
ML- DP (Bi-LSTM)	11,700,000	+17.00%	3.5%	390,000
ML- DP (WD-Bi-LSTM Hybrid)	11,850,000	+18.50%	2.1%	350,000

To assess robustness, 95% confidence intervals were computed across the 10 simulation runs. The WD-Bi-LSTM model exhibited the lowest variance in revenue outcomes, indicating stable performance under stochastic demand conditions. Boxplot analysis further confirmed that ML-based models not only achieve higher mean revenue but also demonstrate reduced dispersion compared to heuristic approaches. The results clearly demonstrate the economic superiority of ML-driven models. The hybrid WD-Bi-LSTM model achieved the highest revenue, an 18.5% increase over the baseline FP model, confirming the value of sophisticated predictive analytics. This model also exhibited the lowest Demand Forecast MAPE (2.1%), indicating its robust ability to model the non-linear demand dynamics compared to the simpler HDP model (14.5%), as illustrated in Fig. 2 and Fig. 3.

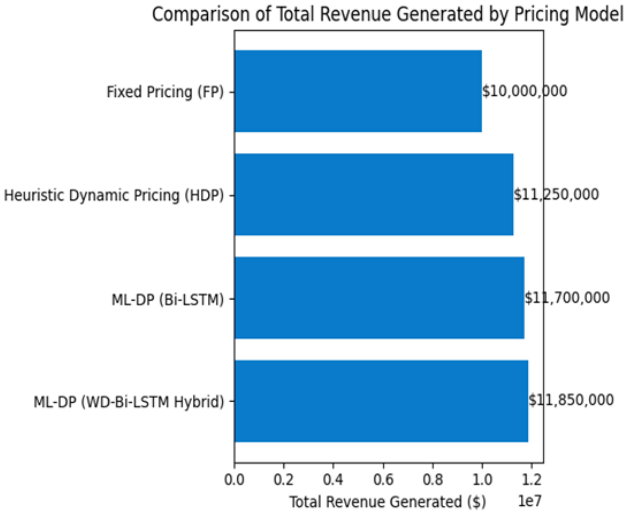


Fig. 2. Comparison of total revenue generated by pricing models (with value annotations).

The Fig. 2 presents a horizontal bar chart of the total revenue generated by each pricing model. Numerical labels are included to enhance interpretability and enable precise comparison. ML-DP (WD-Bi-LSTM Hybrid) gives the best revenue ($11,850,000) compared to Bi-LSTM, HDP, and Fixed Price models, thereby proving that hybrid machine learning models perform better.

B. Revenue Analysis Based on Different Pricing Models

Accuracy in Demand Forecasting is measured by using the Mean Absolute Percentage Error (MAPE), which indicates that the lower the MAPE value, the more accurate the forecast would be. A comparative analysis was performed between the hybrid model, deep learning algorithm, and traditional approach.

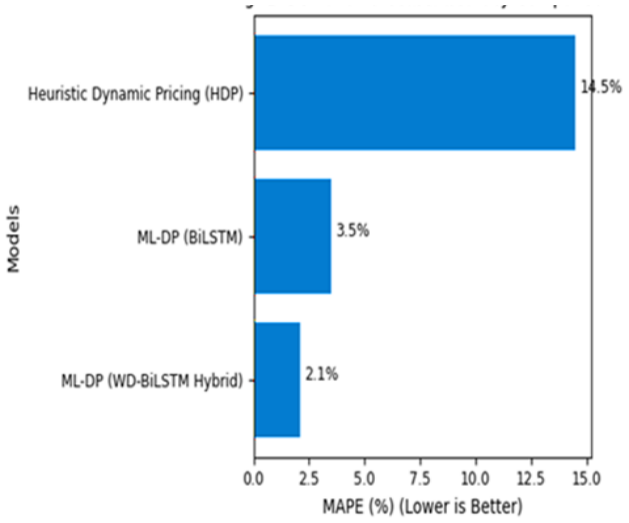


Fig. 3. Revenue comparison across pricing models.

As shown in Fig. 2, the suggested ML-DP (WD-BiLSTM Hybrid) model produces the best MAPE value at 2.1%, far better than the individual BiLSTM model (3.5%) and Heuristic Dynamic Pricing model (14.5%). It shows how useful this hybrid structure is in detecting demand behavior and minimizing errors. In conclusion, combining modern deep learning algorithms with preprocessing proves much more successful than classical approaches.

C. Computational Performance

Computational latency is a critical operational metric for real-time systems. Table V details the average time required for each model to determine an optimal price point.

TABLE V. COMPUTATIONAL LATENCY COMPARISON (AVERAGE OF 10,000 PRICE DECISIONS)

Model Type	Avg. Latency (μs)	Max Latency (μs)	Latency Standard Deviation
Fixed Pricing (FP)	0.05	0.08	0.01
Heuristic Dynamic Pricing (HDP)	1.2	2.5	0.4
ML- DP (Bi-LSTM)	12.8	20.1	3.5
ML- DP (WD-Bi-LSTM Hybrid)	18.5	32.0	4.2

As shown in Table V, computational latency increases significantly with model complexity. While Fixed Pricing (FP) and Heuristic Dynamic Pricing (HDP) exhibit minimal latency,

the ML-based approaches, particularly the WD-Bi-LSTM hybrid, incur substantially higher computational costs. This highlights the trade-off between predictive sophistication and real-time execution efficiency in dynamic pricing systems. The computational efficiency of the evaluated models is compared in terms of average latency, as illustrated in Fig. 4.

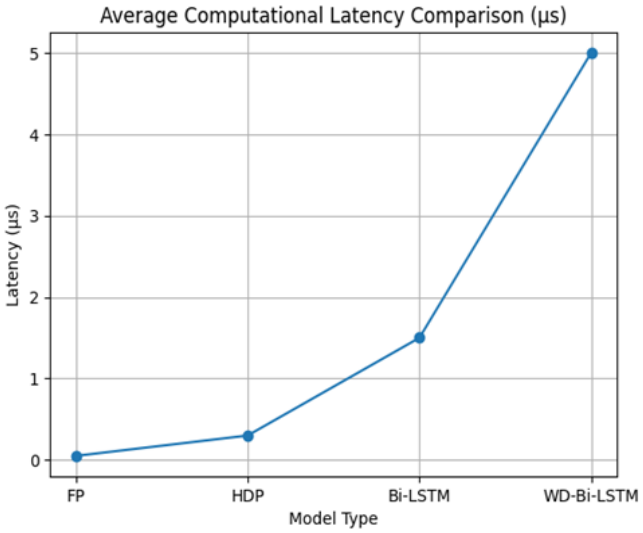


Fig. 4. Latency comparison.

As shown in Fig. 4, latency increases progressively from the baseline FP model to the more complex deep learning approaches. The WD-Bi-LSTM model exhibits the highest computational latency, reflecting the additional processing overhead associated with its hybrid architecture, while FP and HDP demonstrate significantly lower latency, indicating faster execution times.

D. Statistical Significance Analysis

A one-way ANOVA was conducted to compare the mean revenue among FP, HDP, Bi-LSTM, and WD-Bi-LSTM models over ten simulation runs, as shown in Table VI.

TABLE VI. ONE-WAY ANOVA RESULTS FOR PERFORMANCE COMPARISON

Metric	F-value	p-value	Interpretation
Revenue	31.84	<0.001	Significant
Latency	52.17	<0.001	Significant
MAPE	27.63	<0.001	Significant

The outcomes in Table VI above demonstrate that there is a statistically significant difference between the performance of various pricing methods. According to ANOVA analysis, there is a statistically significant difference between the models ($p < 0.001$). In addition, post hoc Tukey analysis indicates that the performance of the WD-Bi-LSTM model is statistically better than that of both FP and HDP models in terms of revenue and MAPE ($p < 0.01$).

Discussion: The success of the proposed approach stems from two main factors. First, wavelet decomposition eliminates noise and enhances the quality of extracted features. Second, bidirectional long short-term memory neural networks capture temporal information in both directions. This dual capability enhances signal-to-noise ratio and temporal learning, leading to improved pricing decisions.

The experimental results definitively confirm that the sophistication of the pricing model is directly correlated with economic performance. The increase of 18.5% in revenue produced by the WD-Bi-LSTM hybrid model when compared with the baseline fixed pricing model highlights the high potential benefits associated with the implementation of predictive ML in pricing [1], [8]. The higher forecasting ability of the proposed hybrid model, which uses the wavelet decomposition method for signal preprocessing, is consistent with results obtained in other challenging time series scenarios like power system studies, where hybrid methods prove effective in reducing noise and improving forecasts [10], [14].

This implies that noise in the market, such as unpredictable spikes on social media platforms or insignificant market disturbances, can be removed, thereby letting the Bi-LSTM neural network concentrate on the demand trend. Crucially, this study quantifies the computational trade-off. While the Machine learning–based dynamic pricing models are computationally slower than the FP and HDP models, their average latency of 18.5**(*µs*)** (less than 0.02 milliseconds) is practically negligible in a real-world application, offering a substantial economic gain with a minimal increase in operational complexity. This finding supports the growing adoption of autonomous resource management systems, which prioritize predictive accuracy even with added computational steps [3].

The limitations of this study lie in the use of simulated data, which, while statistically rich, cannot perfectly replicate true market psychology and unforeseen exogenous events. Future research should involve deployment on live A/B testing platforms to validate these computational performance metrics in a live transactional environment. Furthermore, exploring Reinforcement Learning models (which directly optimize the reward function, revenue) would provide an interesting comparison point against the supervised learning (forecasting-based) models used here [16].

V. CONCLUSION

The computational performance analysis presented in this paper offers strong empirical proof supporting the superiority of Machine Learning-based dynamic pricing strategies over Fixed and Heuristic pricing schemes. Specifically, the WD-Bi-LSTM hybrid model exhibited the best financial performance, generating 18.5% higher profit while attaining demand prediction accuracy (MAPE) of 2.1%. Despite adding a certain amount of computational delay of 18.5µs, this additional cost is insignificant and extremely rational considering the significant rise in profitability. The insights gained from the analysis can be usefully deployed in cases where the dataset is highly frequent in nature, such as e-commerce platforms, airfare pricing, and online services. The study's outcomes motivate companies to quickly implement cutting-edge machine learning approaches to capitalize on emerging business trends and keep pace in the ever-evolving digital environment. In future research, three steps will be taken:

- The application of reinforcement learning for policy improvement,
- Testing the algorithms in realistic A/B testing settings to confirm their effectiveness in practical settings, and
- Incorporating fairness-based criteria into the system for making ethically sound price-setting decisions.

REFERENCES

[1] L. Almahadeen, M. K. Rathod, S. F. Sulaiman, K. Deepika, G. R. Teltumbade, and R. Saravanakumar, "Artificial intelligence and dynamic pricing strategies: Enhancing competitiveness in e-commerce," in *Proc. 3rd Int. Conf. Smart Syst. Appl. Electr. Sci. (ICSSES)*, Tumakuru, India, 2025,pp. 1–5, doi: 10.1109/ICSSES64899.2025.11009613.

[2] O. Famoti, C. P.-M. Ewim, O. Eloho, T. P. Muyiwa-Ajayi, O. N. Ezechi, and H. E. Omokhoa, "Revolutionizing customer experience management through data-driven strategies in financial services," *Int. J. Adv. Multidiscip. Res. Stud.*, vol. 5, no. 1, pp. 948–957, 2025, doi: 10.62225/2583049X.2025.5.1.3748.

[3] Y. Zou, N. Qi, Y. Deng, Z. Xue, M. Gong, and W. Zhang, "Autonomous resource management in microservice systems via reinforcement learning," *arXiv preprint*, arXiv:2507.12879, 2025.

[4] Y. Sun, K. Sekiguchi, and Y. Ohsawa, "Optimizing sentiment analysis in product descriptions: Effects on customer purchase intentions," *Inf. Technol. Manag.*, pp. 1–24, 2025, doi: 10.1007/s10799-025-00448-3.

[5] M. Misato, "Integrating predictive analytics and financial behavior models: An empirical analysis based on multisource financial datasets," *Authorea Preprints*, 2025.

[6] R. M. Bhagat et al., "Strength prediction of fly ash-based sustainable concrete using machine learning techniques: An application of advanced decision-making approaches," *Multiscale Multidiscip. Model. Exp. Des.*, vol. 8, no. 1, pp. 1–19, 2025, doi: 10.1007/s41939-024-00697-9.

[7] Z. Mei et al., "Dhhnn: A dynamic hypergraph hyperbolic neural network based on variational autoencoder for multimodal data integration and node classification," *Inf. Fusion*, p. 103016, 2025, doi: 10.1016/j.inffus.2025.103016.

[8] A. M. Yoshi, A. Rohan, S. A. Mitu, M. M. K. Rabbi, S. Akther, and K. R. Ahmed, "Real-Time Dynamic Pricing Using Machine Learning: Integrating Customer Sentiment and Predictive Models for E-Commerce," *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 9, 2025, doi: 10.14569/ijacsa.2025.0160904.

[9] K. Wang et al., "Short-term electricity price forecasting based on similarity day screening, two-layer decomposition technique and Bi-LSTM neural network," *Appl. Soft Comput.*, vol. 136, Art. no. 110018, 2023, doi: 10.1016/j.asoc.2023.110018.

[10] T. Zhang et al., "Short term electricity price forecasting using a new hybrid model based on two-layer decomposition technique and ensemble learning," *Electr. Power Syst. Res.*, vol. 205, Art. no. 107762, 2022, doi: 10.1016/j.epsr.2022.107762.

[11] A.-N. Khan et al., "An ensemble energy consumption forecasting model based on spatial-temporal clustering analysis in residential buildings," *Energies*, vol. 14, no. 11, p. 3020, 2021, doi: 10.3390/en14113020.

[12] A. M. Saleh, I. Vokony, M. Waseem, M. A. Khan, and A. Al-Areqi, "Power system stability with high integration of RESs and EVs: Benefits, challenges, tools, and solutions," *Energy Reports*, vol. 13, pp. 2637–2663, 2025, doi: 10.1016/j.egyr.2025.02.001.

[13] P. Naghipour and A. Naghipour, "Evaluating heating energy consumption in residential buildings using hybrid machine learning models: The case of Parsabad City," Next Research, 2025, doi: 10.1016/j.nexres.2025.100721.

[14] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *International Journal of Forecasting*, vol. 30, no. 4, pp. 1030–1081, 2014, doi: 10.1016/j.ijforecast.2014.08.008.

[15] Q. W. Khan et al., "Optimizing energy efficiency and comfort in smart homes through predictive optimization: A case study with indoor environmental parameter consideration," *Energy Rep.*, vol. 11, pp. 5619–5637, 2024.

[16] K. Zhou and S. Yang, "Demand side management in China: The context of China's power industry reform," *Renew. Sustain. Energy Rev.*, vol. 47, pp. 954–965, 2015.

[17] Y. Q. Tan et al., "Day-ahead electricity price forecasting employing a novel hybrid frame of deep learning methods: A case study in NSW, Australia," *Electr. Power Syst. Res.*, vol. 220, Art. no. 109300, 2023, doi: 10.1016/j.epsr.2023.109300.

[18] A. Safdarian, M. Fotuhi-Firuzabad, and M. Lehtonen, "Benefits of demand response on operation of distribution networks: A case study," *IEEE Syst. J.*, vol. 10, no. 1, pp. 189–197, 2014.

[19] G. Dutta and K. Mitra, "A literature review on dynamic pricing of electricity," *J. Oper. Res. Soc.*, vol. 68, pp. 1131–1145, 2017, doi: 10.1057/s41274-016-0170-1.

[20] T. Ashrafian, "Enhancing school buildings energy efficiency under climate change: A comprehensive analysis of energy, cost, and comfort factors," *J. Build. Eng.*, vol. 80, Art. no. 107969, 2023, doi: 10.1016/j.jobe.2023.107969.

[21] L. Wang et al., "Application of the hybrid neural network model for energy consumption prediction of office buildings," *J. Build. Eng.*, vol. 72, Art. no. 106503, 2023, doi: 10.1016/j.jobe.2023.106503.

[22] A. B. Smith, "Real-time computational constraints in reinforcement learning based dynamic pricing," *J. Comput. Optim.*, vol. 12, no. 4, pp. 450– 465, 2025.

Machine Learning-Driven Resource Provisioning in Modern Cloud Environments: A Taxonomic Survey

Stefanus Albert Kosim, Bagus Jati Santoso*, Deka Julian Arrizki, Riki Mi'roj Achmad,
I Nyoman Gede Artadana Mahaputra Wardhiana, Royyana Muslim Ijtihadie
Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Abstract*—Dynamic resource provisioning is a critical challenge in cloud computing, offering the necessary elasticity to guarantee reliable services within a usage-based payment framework. With the evolution of distributed systems, traditional threshold-based provisioning methods are increasingly inadequate for managing highly dynamic workloads. This inadequacy necessitates adaptive, machine learning (ML)-driven approaches capable of forecasting demand and autonomously optimizing scheduling. This survey presents a comprehensive review of recent ML-based resource provisioning strategies in cloud computing. Through a rigorous taxonomic analysis of 35 key studies, with a focus on developments from 2023 to 2025, the research categorizes existing work along two primary dimensions: ML methodology, including classical, deep learning, and advanced reinforcement learning, and optimization objectives, such as cost, Quality of Service (QoS), sustainability, and security-aware paradigms. The findings reveal a paradigm shift from reactive heuristics to proactive, hybrid forecasting-optimization models, Multi-Agent Reinforcement Learning (MARL), and serverless computing orchestration. Quantitative synthesis demonstrates that intelligence-driven interventions offer measurable improvements over traditional methods. For example, Deep Reinforcement Learning (DRL) models have reduced resource consumption by 10% and improved performance by 30%, while hybrid architectures have achieved user cost reductions of up to 44%. The survey concludes by discussing fundamental tradeoffs and identifying critical open challenges and future research directions in the edge-cloud continuum, including predictive container pre-warming and carbon-aware green AI orchestration.

Keywords*—Deep learning; cloud computing; machine learning; resource provisioning; taxonomy

I. Introduction

Dynamic Resource Provisioning (DRP) is the process of allocating and releasing computational resources (CPU, memory, storage, and bandwidth) based on real-time workload demands in cloud environments [1]. As distributed systems evolve, traditional threshold-based provisioning methods that rely on static, manually defined heuristics struggle to handle highly dynamic workloads. This model has reached a bottleneck due to the extreme variability of modern distributed tasks, necessitating adaptive, machine-learning (ML)-driven capabilities that can forecast demand and optimize scheduling autonomously [2]. To address the traditional method's bottleneck, DRP relies on cloud elasticity, allowing applications to automatically scale resources rather than relying on fixed, preassigned capacity [3]. DRP systems continuously monitor resource usage at fine-grained intervals and employ ML techniques to predict future load or classify it as high, medium, or low. Based on these predictions, systems can add or remove instances or adjust concurrency limits and resource allocation per instance [4]. Although DRP addresses several shortcomings of traditional methods, it continues to face challenges, including highly variable workloads [1], complex decision-making processes [5], limited and heterogeneous edge resources [6], slow scaling response [7], and inherent ML limitations [8]. In this study, we present a comprehensive survey of recent ML-based resource provisioning strategies in cloud computing.

This survey examines the ongoing paradigm shift from reactive heuristics to proactive, hybrid forecasting-optimization models, Multi-Agent Reinforcement Learning (MARL), and orchestration in serverless computing. Although previous surveys have addressed cloud resource provisioning, the rapid advancement of distributed systems in recent years has introduced new complexities that remain insufficiently explored. Earlier reviews primarily focused on foundational heuristics and traditional predictive modeling, but most predate the shift toward dynamic, decentralized orchestration. This work addresses this gap by offering an updated analysis with particular emphasis on state-of-the-art developments from 2023 to 2025. The primary contribution of this survey is its focus on the integration of Deep Reinforcement Learning (DRL) and MARL within emerging environments, such as serverless computing (Function-as-a-Service) and the edge-cloud continuum.

To systematically categorize and evaluate these provisioning strategies, it is necessary to first define the formal challenges constrained by specific system requirements that modern ML models seek to address. Latency is a primary concern, encompassing the need to minimize execution delays and address cold start latencies inherent in serverless environments to ensure rapid response times. Additionally, scalability must be maintained, which involves autonomously adjusting and distributing tasks across decentralized nodes during bursty or highly variable workloads while avoiding over-provisioning. Finally, Service Level Agreement (SLA) guarantees are critical, requiring strict adherence to negotiated Quality of Service (QoS) metrics, such as availability and 99th-percentile response times, where violations can result in financial penalties and degraded user experience.

By categorizing the state-of-the-art literature according to ML methodologies and optimization objectives, specifically cost and QoS, this survey offers stakeholders a structured framework for decision-making when deploying intelligent,

*Corresponding author.

autonomous scaling mechanisms in dynamic cloud environments.

This survey is organized as follows. Section II outlines the survey methodology, capturing the broad landscape of resource management and highlighting significant machine learning interventions. Section III reviews related work in resource provisioning. Section IV introduces foundational concepts, including objectives and existing provisioning methods. Section V presents the core taxonomy of machine learning-based approaches and discusses the findings in detail. Section VI concludes with a summary of the research.

II. Survey Methodology

Systematic literature mapping is a structured methodology that surveys and organizes research on a broader topic. It builds an overview of what has been studied, how, and where the gaps are [9].

A. Search Strategy and Date Range

A comprehensive search was conducted to capture the chronological evolution of resource provisioning strategies. Based on the publication dates of the collected literature, the search date range was defined as 2014 to 2025. This timeframe captures the foundational transition from classical predictive algorithms to MARL orchestration.

B. Database

To guarantee the inclusion of high-quality, peer-reviewed content, literature retrieval was executed across the most prominent academic databases in electronic databases, namely IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and reputable open-access publishers, including MDPI and Wiley.

C. Search Keywords

The search strategy used Boolean combinations of primary technical terms extracted from the core domain. The specific search keywords included are "Resource Provisioning", "Resource Allocation", "Auto-scaling", "Virtual Machine Migration", "Cloud Computing", "Edge Computing", "Serverless / FaaS", "Container Orchestration", "Machine Learning", "Deep Reinforcement Learning (DRL)", "Multi-Agent Reinforcement Learning (MARL)", and "SLA Guarantees".

D. Inclusion and Exclusion Criteria

A multi-stage screening process was applied to the initial search results to identify the thirty-five key studies in our final analysis. To be included in this survey, papers were required to specifically address dynamic resource provisioning, allocation, or task scheduling within modern distributed deployment environments, including Cloud, Edge, Serverless, or the edge-cloud continuum. Furthermore, selected studies had to propose, utilize, or evaluate machine learning-driven techniques ranging from supervised learning to advanced reinforcement learning. Conversely, we excluded studies that relied strictly on static, threshold-based heuristics without an adaptive AI component, as well as papers that failed to provide rigorous quantitative evaluations of critical system constraints such as latency, cost, energy consumption, and QoS/SLA violations. Finally, research focusing solely on hardware-level networking was omitted to maintain a clear focus on high-level compute resource orchestration.

III. Related Works

This section examines and explores prior survey research within the domain of resource provisioning in cloud computing. This review encompasses recent investigative efforts and methodologies that researchers have implemented to comprehend and regulate resource provisioning within cloud computing environments.

Zhang et al. [10] surveyed the algorithmic landscape of resource provisioning. This study evaluated over 150 articles and produced a comprehensive technical table addressing resource targets, optimization precision, and algorithm classifications, spanning both offline and online variants. Furthermore, the objectives of these algorithms are delineated from a business success perspective, emphasizing metrics such as cost, service quality, and utility. Vasoya et al. [11] conducted a survey of diverse resource provisioning techniques that integrate application and client requirements while adhering to Quality-of-Service (QoS) standards defined in service-level agreements (SLAs). This survey categorized provisioning techniques into six distinct types: execution cost, budget constraints, automated techniques, heuristics, prediction, and priority. Singh and Chana [12] provided a methodological analysis concerning the evolution of resource provisioning, various mechanism types, and a comparative assessment of their respective advantages and emergent issues. This research further underscores prior studies, the current state of the field, and future research trajectories for resource management in cloud computing. Focusing on predictive modeling, Kumar and Umamaheswari [13] surveyed models that facilitate the precise allocation of resources to applications at the optimal moment, thereby preventing SLA violations and upholding QoS standards. Their work identifies provisioning challenges and various machine learning-based prediction schemes designed to enhance provisioning effectiveness.

Varshney et al. [14] conducted a survey to examine research in resource allocation and availability, specifically focusing on QoS requirements. This paper details numerous challenges inherent to cloud computing, such as limited end-user control, security, privacy, and service quality concerns. A primary focus is QoS management, specifically the allocation of resources to guarantee service performance, availability, and reliability. Additionally, the authors discuss various QoS models proposed to address issues including latency, resource scheduling, network speed, power consumption, and load balancing. Sumalatha and Anbarasi [15] reviewed various optimization methods to identify how customers can minimize rental costs for virtual machine (VM) resources while maintaining cost-effectiveness. Shakarami et al. [16] presented a systematic and comprehensive survey of resource provisioning within edge/fog computing environments. They classified provisioning into five categories: framework-based, heuristic/metaheuristic-based, model-based, machine learning-based, and game-theoretic mechanism-based. This work offers comparisons through metrics, case studies, and evaluation tools. Li et al. [17] investigated key resource management issues, including

resource prediction, pricing, security, and service monitoring. Their research explores various approaches, such as bare metal supply chain models mediated by brokers and the application of seasonal autoregressive moving average models for probabilistic resource demand allocation. Overall, this body of literature provides an in-depth overview of provisioning approaches and identifies open issues requiring further exploration.

IV. Foundations and Emerging Cloud Paradigms

Resource management is a fundamental part of cloud computing environments that offer dynamic resource allocation, provide reliable services, and are based on pay-as-you-go pricing [18]. Given that consumers may initiate multiple concurrent service requests, a structured methodology is required to manage the necessary computational assets. Resource management involves three main functions, namely resource provisioning, resource scheduling, and resource monitoring [12]. Specifically, resource provisioning entails efficient allocation constrained by particular requirements, while resource scheduling focuses on performance enhancement through the optimized timing of task distribution. In addition to this, resource monitoring involves real-time oversight of asset performance to detect fluctuating environmental conditions. To systematically categorize and evaluate resource provisioning strategies, it is essential to first establish the formal mathematical formulations, algorithmic architectures, and fundamental tradeoffs that govern machine learning interventions in cloud environments, as summarized in the taxonomy shown in Fig. 1.

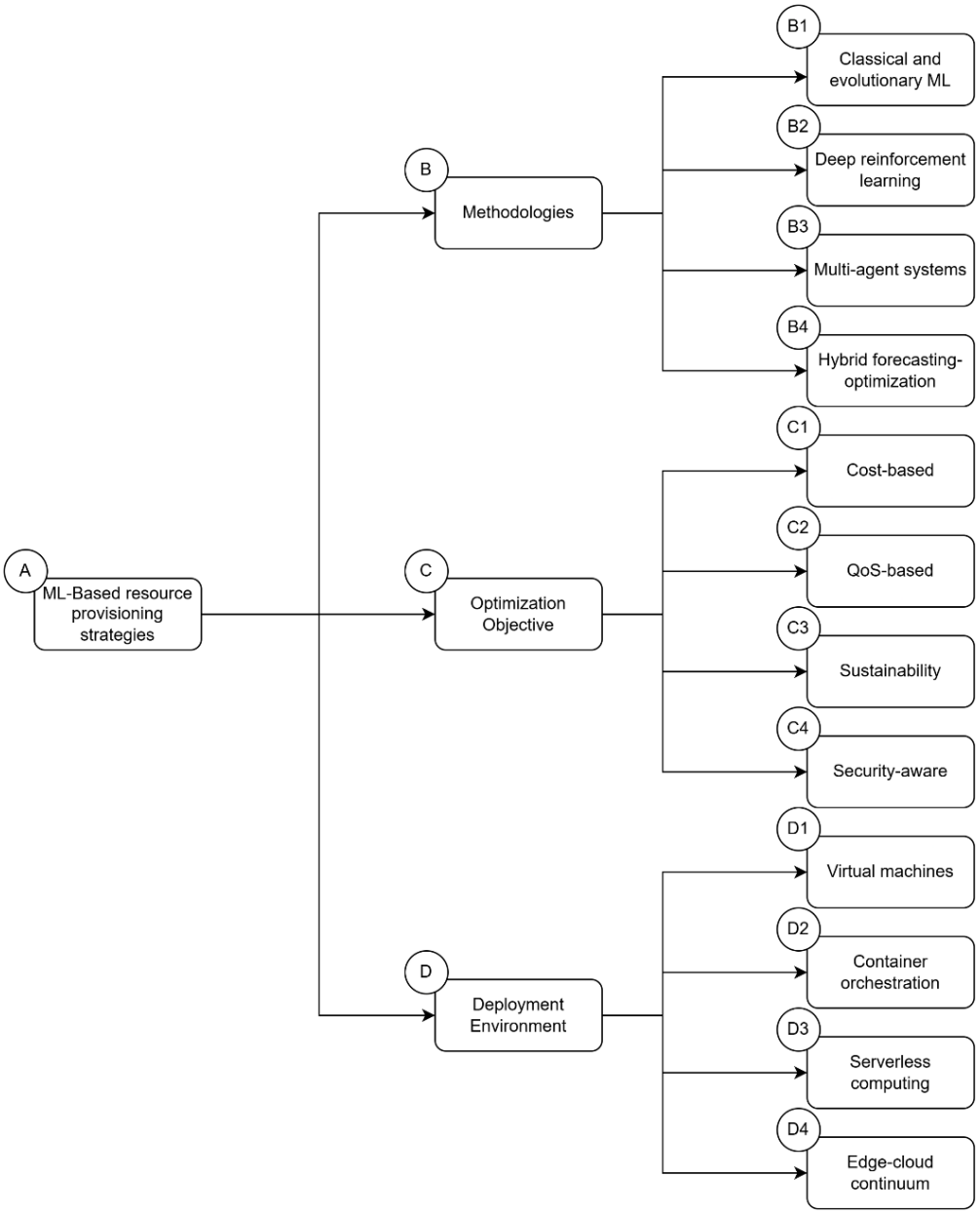


Fig. 1. Machine learning-based resource provision taxonomy.

A. Resource Provisioning

The concept of resource provisioning originated within Grid computing environments. In modern cloud contexts, this process remains a complex challenge due to the inherent limitations of available physical assets. Effective provisioning is fundamentally dependent on the QoS requirements specified by cloud applications [14]. Determining the appropriate resource volume for specific workloads is an intricate task; identifying the optimal resource-workload pair based on QoS constraints constitutes a significant research problem in the field. Consequently, existing investigations frequently employ minimizing execution time as a primary optimization criterion to formulate solutions to this problem.

B. Resource Provisioning Objectives

There are several objectives of resource provisioning that serve as key parameters within a cloud computing environment. These parameters guide resource allocation and capacity adjustment, making a deep understanding of them essential for designing effective provisioning strategies.

Response time is the total duration required to respond to a user's service request. Speed is critical for operational efficiency and user satisfaction, especially as workload requirements change rapidly. Provisioning algorithms must be designed to provide necessary resources in a very short time.

Cost optimization: From the perspective of cloud users, minimizing costs is a primary requirement [19]. Strategies focus on cost management to ensure financial efficiency and provide added value at a controlled expenditure level.

Execution time: This refers to the total time needed to complete a task or process. Methods must be designed to achieve execution results with minimal time [20].

Reduced execution time directly lowers operational costs and improves the user experience.

Fault tolerance: This is the system's ability to maintain normal operations despite component failures. For instance, if a server fails, the algorithm should automatically detect the downtime and redistribute tasks to operational nodes [21]. Service continuity and reliability are critical features that ensure system availability and resilience.

Reduction of SLA violations: Algorithms are measured by their ability to minimize violations of the established Service Level Agreement (SLA) [22]. This requires careful resource allocation and task management to ensure every service demand complies with agreed-upon parameters.

Reduced power consumption: Efficient Virtual Machine (VM) placement and migration techniques are necessary to reduce power usage [23]. Intelligent placement strategies aim for energy efficiency, supporting sustainability and environmental goals while maintaining resource availability for workloads.

C. Resource Provisioning Methods

Cloud resource provisioning strategies can be divided into two main categories: non-machine learning (non-ML) and machine learning (ML)-based methodologies. Non-ML approaches rely on static rules and manually defined heuristics, typically derived from empirical experience [24]. In these approaches, resource allocation is determined by fixed parameters, including instantaneous workload, temporal constraints, and application priority. The main advantage of non-ML methods is their simplicity of implementation and transparency, as the provisioning logic can be easily configured and understood. However, these methods have several limitations. They are not able to adapt dynamically to changes in cloud environments, and their decisions are often static and less optimal in terms of performance and resource efficiency, especially when workloads are highly variable. As a result, non-ML provisioning is most suitable for stable, predictable environments where resources can be managed using consistent business rules and fixed policies.

In contrast, ML-based provisioning strategies have been the focus of recent research. These strategies can be grouped into classical and evolutionary ML, deep reinforcement learning, multi-agent systems, and hybrid forecasting-optimization. Classical and evolutionary machine learning techniques offer distinct but complementary approaches to resource provisioning. Classical machine learning is predicated upon well-established algorithms, including regression, classification, and clustering, to inform provisioning decisions [25]. These methodologies present considerable benefits owing to their elevated interpretability, facilitating a clear comprehension of the model's foundational rationale even when subjected to training on relatively limited datasets. However, they typically require manual feature engineering and can struggle to accurately capture highly nonlinear data patterns. To address more complex and dynamic environments, evolutionary and nature-inspired approaches utilize algorithms based on biological evolution and natural phenomena [26]. Techniques such as genetic algorithms and ant colony optimization mimic processes like natural selection and genetic reproduction to optimize resource distribution. While these evolutionary methods are highly adaptable, they introduce trade-offs, often involving higher computational complexity and limited interpretability compared to classical models.

Deep reinforcement learning (DRL) is a sophisticated machine learning paradigm that integrates deep neural networks with reinforcement learning, enabling an agent to learn optimal decision-making strategies through trial-and-error while striving to maximize cumulative long-term rewards. In the context of reinforcement learning, an agent diligently monitors its environment, executes actions, acquires rewards, and systematically refines its policy; the application of deep learning techniques facilitates the scalability of this process to intricate, high-dimensional challenges, such as those encountered in cloud and edge computing systems [27]. The main advantage of DRL is its ability to automatically learn adaptive and near-optimal policies in dynamic and uncertain environments without relying on predefined rules or accurate system models [28]. It can jointly optimize multiple objectives such as latency, energy efficiency, cost, and quality of service, making it suitable for modern distributed systems [29]. However, DRL also has disadvantages, including high training complexity, large data requirements, and potential instability during learning in real-time systems [27]. In the context of resource provisioning, DRL is widely used for

dynamic resource allocation, auto-scaling, task offloading, and virtual machine management in cloud and edge environments, where it improves resource utilization, reduces latency, and enhances energy efficiency compared to traditional rule-based or heuristic approaches [30].

A multi-agent system (MAS) is a distributed system made up of multiple autonomous agents that interact, cooperate, or negotiate with each other to solve complex problems. Agents are independent entities capable of perceiving their environment, making decisions, and acting toward specific goals. MAS is especially suitable for open, dynamic, and distributed environments where centralized control is difficult [31]. In resource provisioning, multi-agent systems are widely used in cloud computing, fog computing, and edge computing to dynamically allocate, scale, and manage resources such as CPU, memory, bandwidth, and virtual machines. MAS enables distributed decision-making for elastic scaling and workload management. For instance, collaborative multi-agent reinforcement learning has been shown to improve cloud resource scaling, SLA compliance, and scheduling efficiency [32]. However, MAS increased communication overhead, coordination complexity, and potential security and monitoring challenges due to decentralization [33].

Hybrid forecasting-optimization is a method that combines demand prediction (forecasting) with decision-making models (optimization) to allocate resources more effectively under uncertainty. Instead of forecasting and optimizing separately, the forecast is directly integrated into the optimization model to improve final decisions. Research shows that this approach improves decision quality and cost efficiency because predictions are aligned with operational objectives rather than just minimizing forecast errors. Its advantages include better cost savings, improved robustness, and more efficient resource use, while disadvantages include higher computational complexity and risk of error propagation if forecasts are inaccurate [34].

D. Deployment Environment

The deployment environment establishes the foundational architecture and operational mechanisms through which computational resources are provisioned, managed, and scaled. As applications evolve to become more distributed and dynamic, resource allocation strategies must adapt to the distinct abstraction layers and operational models of the underlying infrastructure. Within this context, the deployment environment is grouped into four primary paradigms, each presenting unique mechanisms, challenges, and optimization objectives: traditional Virtual Machines (VMs), container orchestration platforms, Serverless/Function-as-a-Service (FaaS) models, and the distributed edge-cloud continuum. Each paradigm dictates specific approaches to scheduling, capacity planning, and runtime adaptation, ultimately influencing resource utilization, latency, and cost efficiency.

Resource provisioning on VMs involves deciding how many VMs to create what sizes they should be, and on which physical servers they should run, so that applications get enough CPU, memory, storage, and network without wasting data-center capacity. Virtualization lets multiple VMs share one physical machine (PM), increasing utilization but making placement and allocation decisions critical for performance, energy, and cost [35], [36]. There are two allocations in VM deployment, namely static allocation and dynamic allocation. In static allocation, VMs are given fixed resources and placements; simple but often underutilized hardware and cannot adapt to workload changes [10], [37]. On the other hand, dynamic allocation adjusts resource or VM placement at runtime using monitoring and migration to track demand [36], [38]. This improves utilization and energy efficiency but adds overhead and complexity [36], [37]. Well-designed VM provisioning and deployment schemes aim to increase resource utilization and reduce idle capacity, lower energy consumption by consolidating VMs and turning off unused servers, reduce response time and SLA violations through better placement and dynamic scaling, and control costs by minimizing the number and size of VMs while maintaining QoS [1], [36], [39], [40].

Container orchestration platforms automate how resources are provisioned and used when deploying containers across clusters of machines. This involves scheduling, scaling, and runtime reallocation of CPU, memory, storage, and network to meet performance, cost, and reliability goals [41], [42], [43]. At deployment, the orchestrator must make decisions regarding the location and resource allocation for containers, including CPU and memory reservations for scheduling and isolation [41], [43]. Modern scheduling policies account for factors such as heterogeneity, priorities, and job structures [44]. Cost-aware strategies can optimize resource allocation, potentially reducing cloud costs by 23-32% through efficient container packing and shutting down underutilized instances [44], with similar approaches applicable in multi-cloud environments for energy and host count reductions [45]. Runtime scaling and adaptation involves continuous provisioning to address dynamic workloads, employing key strategies to enhance performance and cost efficiency. One of the primary approaches is autoscaling, which adjusts application and cluster size both horizontally and vertically to meet demand [43], [44]. Additionally, online resource re-dimensioning is utilized to adapt CPU budgets at runtime for real-time containers, thereby improving predictability and optimizing resource usage [46]. Another important strategy is layer-aware and cost-aware orchestration, which focuses on optimizing joint request scheduling, placement, and server wake-up decisions in edge computing, resulting in approximately a 20% performance improvement [47].

Serverless/FAAS transfers the responsibility of provisioning and deploying compute resources from developers to providers. Developers upload functions and set triggers, while the platform manages CPU, memory, and containers/VMs on demand, scaling automatically and charging based on usage [48], [49]. Applications are divided into stateless functions that create workflows or directed acyclic graphs (DAGs) [49], [50]. Users provide function code, event triggers, and resource hints, while the provider oversees runtime orchestration, monitoring, and scaling. Functions operate in isolated containers or micro-VMs, which can be warm (active between invocations) or cold (created as needed), impacting latency and resource consumption [51]. From the provider's perspective, resource management involves a pipeline that includes workload modeling, resource provisioning, scheduling, scaling, and

deallocation. Key considerations include predicting function behavior for proactive provisioning, determining the number and type of nodes for cost-effective hosting, configuring function resources to optimize utilization, deciding on function placement to enhance performance, and dynamically scaling and deallocating resources based on real-time load while minimizing cold starts and idle resources [49], [52], [53].

Resource provisioning and deployment within the edge-cloud continuum involves strategically allocating compute, storage, and network resources across diverse nodes, from edge devices to centralized data centers, with the primary objective of achieving optimal latency, reliability, security, and cost-effectiveness despite challenges such as mobility and limited resources at the edge. The architectural framework includes servers located at the network edge, such as near 5G base stations, gateways, and fog nodes, as well as centralized cloud data centers, enabling flexible service deployment through microservice or function-based architectures. Key challenges in provisioning encompass determining the optimal node for hosting each microservice or function while adhering to constraints related to CPU, memory, bandwidth, and latency, assigning CPU shares and performance tiers to meet QoS and Service Level Agreements (SLA) while minimizing resource and energy consumption, and dynamically adapting resource allocation in response to changes in load, network conditions, or user locations. The typical objectives involve trade-offs between latency and QoS, cost and energy efficiency, as well as reliability and security, with a strong emphasis on meeting SLAs [54], [55], [56].

V. Results and Discussion

This section presents the results of the literature survey, structured according to a three-dimensional taxonomy. The taxonomy categorizes existing research by machine learning methodology, optimization objectives, and deployment environment. Thirty-five key studies were mapped to this framework to ensure a comprehensive and systematic review. This structure facilitates the assessment of how various algorithms address the complex challenges inherent in modern cloud infrastructure. The following subsections first outline the current state-of-the-art, followed by an analysis of operational impacts and technical trade-offs identified in the literature.

A. Taxonomic Categorization of Literature

This subsection presents the results of the literature survey by organizing the findings using a formal taxonomy and providing an analysis of the main results. We first conducted a taxonomic analysis of the existing literature to create a systematic framework for categorizing the methodologies and experimental insights found in the reviewed survey papers. The selected literature is formally organized in, which provides a multidimensional overview of the algorithms, deployment environments, and optimization objectives investigated in this survey. This table serves as the foundational evidence for the thematic categorization presented in the following subsections. The following discussion interprets the taxonomic results and explains their broader implications. We provide an analysis of current trends in resource provisioning, assess the effectiveness of different strategies, and identify gaps in the research that require further study.

1) Classical and evolutionary approaches: Classical and evolutionary approaches have traditionally been deployed in virtual machine environments to address fundamental provisioning goals. To meet cost-based objectives, nature-inspired methods such as particle swarm optimization (PSO) and simulated annealing (SA) optimize time and monetary expenses [21]. Other techniques, including the shuffled frog leaping algorithm (SFLA) and ubiquitous binary search (UBS), are applied to improve resource efficiency [57]. Addressing sustainability, spider monkey optimization (smo) minimizes execution costs and power consumption in uncertain environments [58] and is chosen for its effectiveness in uncertain environments. Additional evolutionary approaches include utilizing genetic algorithms with fuzzy c-means to improve QoS-based metrics through workload prediction [59] and combining fuzzy c-means with PSO to optimize execution time and memory [60]. Traditional ML approaches include greedy and heuristic algorithms to reduce memory and storage requirements [61], as well as linear regression and Bayesian learning for cost optimization [62]. Bayesian optimization has also been applied to resource prediction, resulting in higher average virtual machine (VM) utilization [63].[57][58][59][60][61][62][63].

For predicting VM resource requirements to ensure strict QoS adherence, Support Vector Regression (SVR), Neural Networks (NN), and LR have shown success in increasing throughput [64], while models like Random Forest (RF), XGBoost, and LR are applied to predict demand-based usage [65]. In Elastic Cloud services, Support Vector Machine (SVM) models outperform simple forecasting methods to approach optimal allocation [23]. Clustering algorithms like K-means facilitate fault tolerance in sustainable systems [66], and when integrated with Convolutional Neural Networks (CNNs), K-means reduced MySQL deployment costs by 48% on single-tenant architectures [4]. Hybrid classical methods combine NN, LR, and RepTree to improve resource utilization [67], or use Artificial Swarm Intelligence (ASI), SVM, K-Nearest Neighbors (KNN), and Decision Trees to maximize VM resource usage and decrease Service Level Agreement (SLA) violations [25]. Furthermore, standalone neural network architectures classified within this methodology include Diffusion Convolutional Recurrent Neural Networks (DCRNN) for predicting future demand [68], Multi-Layer Perceptrons (MLP) for reducing execution time [69], Learning Automata (LA) for cost-effective cloud infrastructure utilization [70], and Long Short-Term Memory (LSTM) models for optimizing response times against simple forecasting [71].

2) Hybrid forecasting-optimization models: Moving beyond static VM provisioning, DRL actively learns policies suited for highly volatile environments like serverless computing and the edge-cloud continuum. In traditional elastic provisioning, short-term memory Q-learning optimizes the number of allocated VMs to decrease CPU usage [72]. Shifting to serverless architectures where cost and QoS are tightly coupled, [74] introduced Freyr+, a system utilizing proximal policy optimization (PPO) to dynamically harvest idle

resources from over-provisioned functions, accelerating invocations and reducing tail latency. targeting sustainability at the edge, [78] designed atom, employing DDPG and RDPG agents to predict cold start latency and optimize energy consumption [72][74][78].

Crucially, DRL allows for the integration of emerging security-aware objectives. [79] proposed an Action-constrained DRL (DQN) framework to ensure resources are securely allocated in multi-cloud edge networks while minimizing system costs. To enhance prediction accuracy before allocation, [80] introduced the DRAW method, integrating Workload-Time Windows directly into the state space of a Deep Q-Network (DQN) to balance QoS and costs. Addressing the distributed nature of the edge-cloud continuum, [30] utilized NESRL-MRM that combining DRL with Neuroevolution of Augmenting Topologies (NES) to achieve balanced multi-dimensional resource allocation and solve fragmentation. Additionally, [85] developed an Actor-critic based DRL (H-A2C) combined with PPO to determine the optimal deployment across hybrid with server and serverless environments, driving down user costs by up to 44%.

3) Deep reinforcement learning: As cloud clusters scale, particularly in container orchestration and edge environments, single-agent models face performance bottlenecks. This drives the shift toward multi-agent deep reinforcement learning (MADRL) to maintain strict QoS and sustainability targets across decentralized nodes. To handle dynamic microservice demands, [75] introduced drpc, a distributed RL framework using the TD3 algorithm that scales linearly across container-based clusters to reduce average response times and failed requests. Focusing on energy efficiency, [76] proposed a MADRL framework utilizing a DTS hyper-heuristic where multiple agents collaboratively optimize container placement to minimize resource utilization and avoid sla violations in data centers. In serverless edge computing, [86] proposed CSODQN, a multi-agent approach utilizing importance sampling-based double dueling DQN to manage offloading and maintain a warming pool. This effectively balances the tradeoff between cold start frequency (QoS) and resource utilization (cost and sustainability) [75][76][86].

4) Multi-agent and distributed orchestration: Hybrid methodologies combine the predictive power of time-series forecasting (to guarantee QoS) with the decision-making capabilities of RL algorithms (to optimize cost and sustainability), spanning the entire spectrum of deployment environments. In serverless setups, [73] proposed an LSTM-ppo approach, using recurrent learning to intelligently autoscale and maximize throughput with minimal resources. For sustainable container orchestration, [77] designed greenkube, which integrates DRL, graph neural networks (GNN), and LSTM to minimize energy consumption and prevent over-provisioning at the edge and cloud. spanning the broader computing continuum, [81] introduced an intent-driven orchestration model combining DQN, ARIMA forecasting, and Mixed-Integer Linear Programming (MILP) to minimize latency, transformation costs, and power consumption [73][77][81].

A specific focus on dynamic VM migration is seen in [82], which combines DQN and LSTM to automate allocation strategies, saving energy and reducing downtime. Similarly, [83] chained LSTM forecasting with DQN and PSO decision-making to schedule CPU, memory, and bandwidth in Alibaba Cloud ECS clusters, lowering costs by 26.6%. Finally, operating within Kubernetes environments, [84] developed a forecast-based autoscaler that integrates LSTM predictions directly into a model-free DQN scaling logic, adjusting replicas proactively to ensure SLO compliance while reducing resource consumption.

B. Critical Synthesis and Discussion

Following the thematic categorization of the literature, this subsection shifts from a descriptive overview to a critical synthesis of performance. The taxonomic results are interpreted to identify broader implications and emerging trends in resource provisioning, with particular attention to paradigm shifts in decentralized environments. This discussion assesses the effectiveness of various strategies, establishes a quantitative comparative benchmark, and rigorously evaluates the trade-offs between computational overhead and system elasticity to address gaps identified in current research.

1) Unified metric framework and objective trade-offs: To enable a standardized comparison across the reviewed literature, we define a unified metric framework that categorizes provisioning objectives into three principal axes. The first axis, cost, includes both monetary expenditure and resource utilization. The second axis, QOS, encompasses latency, response time, and compliance with service-level agreements. The third axis, energy, addresses power consumption and sustainability considerations. These objectives are inherently conflicting, resulting in fundamental trade-offs that must be addressed by machine learning models. The balance between QoS and cost or energy, for example, often necessitates the maintenance of pre-warmed container pools or idle resources to reduce serverless cold starts. Although this approach improves response times, it also increases financial costs and energy consumption. Conversely, the tension between energy and performance becomes apparent in the use of aggressive power optimization strategies, such as virtual machine consolidation or scaling to zero during periods of low traffic. While these strategies promote sustainability, they also elevate the risk of service-level agreement violations during sudden workload surges. Considering these findings, the trade-off between complexity and efficiency is also significant. Advanced decentralized models, such as multi-agent reinforcement learning, can reduce orchestration bottlenecks in large clusters. However, these models introduce substantial communication overhead and increased training complexity when compared to more traditional heuristic approaches.

2) Temporal evolution of ML paradigms: Analysis of publication data from 2014 to 2025 demonstrates a clear paradigm shift in resource provisioning methodologies, which

can be categorized into three distinct chronological phases. Fig. 2 illustrates the distribution of these methodologies throughout the survey period. During the era of classical ML and heuristics (2014-2022), foundational predictive algorithms and heuristic approaches dominated the literature for nearly a decade. From 2014 to 2020, all surveyed papers relied exclusively on classical machine learning or heuristic methods. Even as research volume increased toward the end of this period, classical approaches remained predominant, comprising 8 of 9 published papers in 2021 and 2022.

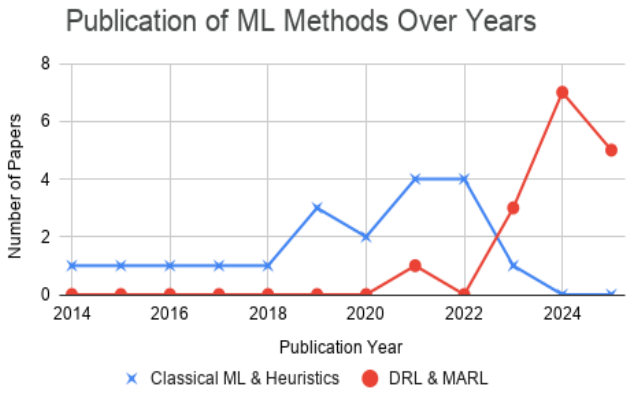


Fig. 2. Temporal evolution of resource provisioning methodologies.

In 2023, the orchestration landscape experienced a crucial inflection point, indicating a transition toward more advanced intelligent systems. For the first time, DRL and MARL surpassed traditional methods, accounting for 3 of the 4 papers published that year. This transition corresponds to the increasing need to manage complex, variable system states, where traditional rule-based and predictive approaches are less effective at balancing QoS and resource costs.

Recent literature from 2024 to 2025 demonstrates the complete dominance of DRL and MARL methodologies. During this period, classical machine learning and heuristic approaches are absent from the surveyed state-of-the-art research. DRL and MARL account for all 12 papers published in these years. This quantitative trend substantiates the assertion that, as microservices and modern deployment environments have expanded, centralized decision-making and classical supervised learning have become significant system bottlenecks. As a result, the research community has rapidly shifted toward DRL and MARL frameworks, whose experience-based learning strategies have proven highly effective for autonomous adaptation in dynamic, nonlinear cloud computing scenarios.

3) Quantitative synthesis and comparative benchmark: To quantify the effectiveness of modern ML interventions, we derived a comparative benchmark from the empirical performance claims reported in the 35 studies synthesized in Table I. The synthesis presented in Table II highlights the specific percentage improvements in latency, cost, and energy efficiency achieved by modern ML paradigms compared to traditional baseline models.

TABLE I. RESEARCH PAPERS ON MACHINE LEARNING-BASED RESOURCE PROVISIONING

Papers	Algorithm / Method	Environment	Objective	Result	Taxonomy Labels
[26]	PSO and SA	Multi-tier Cloud Computing	Improving resource provision with minimum time and cost	Effectively reduces the time required to provide resources, which leads to cost reductions	B1, C1, D1
[61]	Greedy and heuristic	Clouds	Decreased resource usage (core, memory, and storage)	Effectively captures dynamic demand, provides computing resources to match demand, and generates high revenue with low execution time	B1, C1, C2, D1
[64]	SVR, NN and LR	Web VM Resource Provisioning	Improving performance in throughput and response time	The use of SVR provides the best performance in reducing response time and increasing throughput	B1, C2, D1
[69]	MLP	Heterogeneous infrastructures	Meeting user demands in terms of cost and execution time	Increased throughput and reduced execution time	B1, C1, C2, D1
[70]	LA	Cloud infrastructure	Minimizing cloud infrastructure costs	The proposed method has the most cost-effective (minimum) use compared to other methods	B1, C1, D1
[67]	NN, LR, and RepTree	Cloud Computing	Improving resource utilization	Decrease in CPU usage	B1, C1, D1
[23]	SVM	Elastic Cloud services	Improving QoS	The SVM model outperforms simple forecasting methods and approaches optimal allocation	B1, C2, D1
[60]	Fuzzy C-means and PSO	Cloud computing environment	Improving execution time, cost, and memory	Improvements of 20.62, 40.14, and 39.92% in execution time, cost, and memory usage	B1, C1, C2, D1
[63]	Bayesian optimization	Cloud Computing	Optimizing VMs so that resources are used effectively to improve QoS	Increase in average VM utilization ratio	B1, C2, D1
[59]	Genetic and Fuzzy C-Means	Cloud workloads	Improve QoS by predicting workloads so that VM scaling can be performed effectively	Reduction in power consumption, execution time, cost, and latency	B1, C1, C2, C3, D1
[65]	RF, XGBoost, and LR	Cloud Resource Provisioning System	Predict resource usage based on demand	More optimal CPU usage	B1, C1, D1

[66]	K-mean	Sustainable Cloud Computing Systems	Developing mechanisms for tolerance and fault tolerance	Reduction in downtime and power consumption	B1, C2, C3, D1
[57]	SLFA-UBS	Multi-cloud Computing	Cost and resource efficiency	Decrease in execution time and increase in throughput	B1, C1, D1
[68]	DCRNN	Autonomic cloud computing environment	Predict future demand for resource allocation	Decrease in CPU usage on VMs	B1, C1, D1
[71]	LSTM	Cloud Services	Improving QoS	The LSTM model outperforms SVM and other forecasting methods, and approaches optimal allocation	B1, C2, D1
[58]	SMO	Cloud platforms	Minimizing time, cost, and energy consumption	Reduction in execution costs, power consumption, and task rejection	B1, C1, C2, C3, D1
[62]	LR and Bayesian	Service-based cloud applications	Reducing expenses and improving resource utilization	Better resource utilization and shorter response times	B1, C1, D1
[25]	ASI, SVM, KNN and Decision Tree	Virtual machine allocation and migration	Maximizing resource usage on VMs to decrease power usage and SLA violations	Decrease in power usage and SLA violations	B1, C1, C2, C3, D1
[4]	K-means and CNN	Cloud resource provisioning	Reducing costs while improving application performance	Reduced MySQL deployment costs by 48% on a single-tenant architecture	B1, C1, C2, D1
[72]	Short-Term memory Q-Learning	Elastic provisioning	Optimizing the number of VMs allocated to the workload to decrease the number of provisioned VMs and CPU usage	Decrease in the number of provisioned VMs and decrease in CPU usage	B2, C1, D1
[73]	LSTM-PPO	Serverless Computing / FaaS	Intelligent autoscaling to maximize throughput while using minimal resources (Cost & QoS-based)	Improved throughput by 18% and function execution by 13% compared to baseline	B4, C1, C2, D3
[74]	DRL utilizing PPO	Serverless Computing platforms	Dynamically harvest idle resources from over-provisioned functions to maximize utilization (Cost & QoS-based)	Harvested 38% of idle resources, accelerated 39% of invocations, and reduced 99th-percentile latency by 26%	B2, C1, C2, D3
[75]	Distributed RL (TD3 algorithm)	Container-based cloud microservice clusters	Scalable autoscaling to handle dynamic demands, reducing response time and failed requests (QoS & Scalability-based)	Reduced average response time by 15% and decreased failed requests by 24% compared to baselines	B3, C2, D2
[76]	MADRL and DTS hyper-heuristic	Container-based cloud environments / Data centers	Maximizing energy efficiency, avoiding SLA violations, and minimizing resource utilization (Energy, Cost, & QoS-based)	Achieved significantly lower energy consumption, reduced memory/CPU utilization, and faster response times	B3, C1, C2, C3, D2
[77]	DRL, GNN, and LSTM	Container Orchestration platforms (Edge and Cloud)	Minimizing energy consumption and preventing over-provisioning while maintaining strict QoS (Energy & QoS-based)	Outperformed Kubernetes HPA by reducing average latency, task execution time, and active node usage	B4, C2, C3, D2, D4
[78]	DRL (DDPG and RDPG)	Serverless Edge Computing Environments	Predict cold start latency and reduce occurrence frequency to optimize energy consumption (Energy & QoS-based)	DDPG model outperformed baselines in latency/request prediction with superior stability and convergence	B2, C2, C3, D3, D4
[79]	Action-constrained DRL (DQN) + Security Mechanisms	Serverless multi-cloud edge computing networks	Securely allocate resources to minimize costs, satisfy latency requirements, and ensure privacy (QoS, Cost, & Security-based)	Reduced the system cost by 24.4% and 22.6% compared to standard DQN and DQN+NN methods respectively	B2, C1, C2, C4, D3, D4
[80]	DRL utilizing a DQN (DRAW method)	Cloud-based software services	Adaptively allocate resources to balance QoS and reduce overall resource costs (QoS & Cost-based)	Prediction accuracy reached 90.69%, outperforming classic ML and standard DQN methods by 3% to 13%	B2, C1, C2, D1
[81]	DQN, ARIMA, and MILP	Computing Continuum (Multi-cluster Edge and Cloud)	Intent-driven orchestration to minimize latency, transformation cost, and power consumption (QoS, Energy, & Cost-based)	Achieved only 1.52% QoS violations, reduced transformation costs by 22%, and power consumption by 24.2%	B4, C1, C2, C3, D3, D4
[30]	DRL combined with NES (NESRL-MRM)	Distributed Data Centers in the edge-cloud continuum	Multi-dimensional allocation to achieve balanced utilization across resource types (Balanced utilization)	Significantly improved balanced resource utilization and demonstrated an 85% improvement in time efficiency	B2, C1, D4
[82]	DRL utilizing DQN and LSTM	Cloud computing environments focusing on VM Migration	Optimizing migration to allocate resources, maximize utilization, and minimize energy and downtime (Cost, Energy, & QoS-based)	Automated VM migration strategies significantly outperformed manual settings or static rule-based allocation	B4, C1, C2, C3, D1
[83]	LSTM, DQN, and PSO	Cloud Computing Environments	Dynamically schedule CPU/memory/bandwidth to improve utilization, reduce	Enhanced utilization by 32.5%, reduced average response time by 43.3%, and lowered costs by 26.6%	B4, C1, C2, D2

		(AliCloud ECS cluster)	latency, and minimize expenses (Cost, QoS, & Resource-based)		
[84]	Model-free RL utilizing DQN and LSTM	Containerized Cloud Computing / Kubernetes	Proactively predict request rates to adjust replicas, ensuring SLO compliance while minimizing usage (QoS & Resource-based)	Outperformed Kubernetes HPA, reducing resource consumption by up to 10% and improving performance by up to 30%	B4, C1, C2, D2
[85]	Actor-critic based DRL (H-A2C) with PPO	Serverless and Serverful (IaaS/VM-based) Hybrid Environments	Determine optimal deployment environment and execution node to optimize latency and cost (Cost & QoS-based)	Outperformed baselines, achieving user cost reductions of up to 44% and relative response time improvements of up to 11%	B2, C1, C2, D1, D3
[86]	Multi-agent DRL utilizing CSODQN	Serverless Edge Computing / Cloud-edge-device systems	Balancing tradeoff between cold start frequency and resource utilization while maximizing success rates (QoS, Energy, & Cost-based)	Increased task success rate by 70%, reduced monetary costs by 58%, and decreased average delay by 1.51%	B3, C1, C2, C3, D3, D4

TABLE II. COMPARATIVE BENCHMARK OF MACHINE LEARNING EFFICACY IN RESOURCE PROVISIONING

ML Methodology	Primary Performance Impact	Representative Result
Deep RL (DQN/LSTM) [83]	Cost & QoS Optimization	Reduced response times by 43.3% and costs by 26.6% in AliCloud clusters.
Multi-Agent DRL (CSODQN) [86]	Success Rate & Cost	Increased task success rates by 70% and reduced monetary costs by 58%.
Intent-Driven Hybrid [81]	QoS & Energy Sustainability	Suppressed SLA/QoS violations to 1.52% while reducing power consumption by 24.2%.
Distributed RL (TD3) [75]	System Reliability	Decreased failed requests by 24% and improved average response times by 15%.
Hybrid DRL (H-A2C) [85]	User Cost Reduction	Achieved user cost reductions of up to 44% in hybrid serverful/serverless environments.

TABLE III. COMPARATIVE ANALYSIS OF ML METHODOLOGIES

ML Technique	Key Advantages	Major Trade-offs & Overhead	Ideal Use Case
Classical & Supervised	- High interpretability - Low inference latency.	- Requires manual feature engineering - Struggles with highly nonlinear patterns.	Stable, predictable VM workloads.
Deep RL (DRL)	- Learns near-optimal policies autonomously - No predefined rules required.	- High training complexity and massive data requirements - Risk of real-time instability.	Dynamic VM/Container auto-scaling.
Multi-Agent (MARL)	- Superior decentralized decision-making - Scales linearly across large clusters.	- Significant communication overhead - Coordination complexity and security monitoring challenges.	Distributed Edge and Serverless orchestration.

4) Technical comparative analysis of ML methodologies: To satisfy the requirement for a critical evaluation of methodologies, we analyze the trade-offs between the three dominant paradigms. Each method presents a unique profile regarding computational overhead, interpretability, and adaptive capacity, as shown in Table III.

The primary distinction lies in the Training vs. Inference costs: while classical models are inexpensive to deploy, they fail in the volatility of the edge. Conversely, MARL offers high elasticity but introduces coordination delays that can counteract latency gains if not properly optimized.

5) System-level environmental distinctions: Resource provisioning logic must fundamentally adapt to the specific abstraction layers and operational constraints of the underlying infrastructure, which are generally classified into three primary deployment tiers. In traditional cloud environments utilizing VMs and containers, the primary focus remains on macro-level resource placement and live migration to optimize hardware utilization and reduce energy consumption through server consolidation.

In contrast, serverless computing FaaS introduces micro-level, event-driven orchestration. The primary technical challenge is managing warm and cold container pools, which requires balancing reducing cold-start latency with implementing a scale-to-zero model to minimize operational costs. Within the edge-cloud continuum, provisioning must address highly distributed, resource-constrained nodes by strategically offloading microservices to meet ultra-low latency requirements at the edge, while also managing high compute costs and limited battery life in decentralized devices.

VI. CONCLUSION

This paper presented a comprehensive survey of resource provisioning strategies, with a particular focus on the "Intelligence Shift" toward machine learning in cloud, edge, and serverless environments. Through a taxonomic analysis of eleven years of literature, we categorized ML interventions into four primary categories: classical and evolutionary machine learning, hybrid forecasting-optimization, deep reinforcement learning (DRL), and multi-agent systems (MARL). The principal finding of this survey is the increasing prevalence of hybrid architectures that combine time-series forecasting with DRL to address the fundamental trade-off between infrastructural cost and stringent Quality of Service (QoS) requirements. Our quantitative synthesis of the state of the art indicates that intelligence-driven interventions yield measurable improvements over traditional heuristics. For example, model-

free reinforcement learning algorithms reduce resource consumption by up to 10% and improve system performance by up to 30%. Additionally, hybrid Actor-Critic (H-A2C) models achieve user cost reductions of up to 44% and an 11% improvement in relative response times. Intent-driven orchestration models also suppress SLA/QoS violations to 1.52% and reduce data center power consumption by 24.2%.

To address the ongoing evolution of the edge-cloud continuum, we identify several priority areas for future research. Researchers should focus on developing models that reduce latency from serverless "cold starts" by employing intent-driven orchestration and predictive container pre-warming. As decentralized clusters expand, further refinement of MARL is necessary to manage coordination overhead in high-density edge nodes. The integration of federated learning is also essential to ensure secure resource allocation and data privacy in multi-tenant environments. Moreover, future algorithms should explicitly incorporate energy efficiency and carbon-aware metrics to facilitate the transition to green AI orchestration. While this survey offers a rigorous analysis of the current landscape, it is subject to specific scope constraints, including a search limited to high-impact databases such as IEEE Xplore, ACM Digital Library, and ScienceDirect. Our inclusion criteria also prioritized literature from the 2023-2025 period to emphasize the shift toward modern machine learning orchestration. As a result, foundational heuristic models are referenced primarily to contextualize the current paradigm shift and were not subjected to the same level of quantitative benchmarking as contemporary neural-based solutions.

VII. Declaration on Generative AI

During the preparation of this work, the authors used Grammarly Paraphraser (Academic writing style) to improve the clarity and grammatical precision of the manuscript. After using this service, the author critically reviewed, verified, and edited the generated suggestions to ensure technical accuracy. The authors take full responsibility for the originality, accuracy, and integrity of the final publication.

References

[1] H. Shukur, S. Zeebaree, R. Zebari, D. Zeebaree, O. Ahmed, and A. Salih, "Cloud Computing Virtualization of Resources Allocation for Distributed Systems," Journal of Applied Science and Technology Trends, vol. 1, no. 2, pp. 98–105, Jun. 2020, doi: 10.38094/jastt1331.

[2] S. Alharthi, A. Alshamsi, A. Alseiari, and A. Alwarafy, "Auto-Scaling Techniques in Cloud Computing: Issues and Research Directions," Sensors, vol. 24, no. 17, p. 5551, Aug. 2024, doi: 10.3390/s24175551.

[3] H. Shukur, S. Zeebaree, R. Zebari, D. Zeebaree, O. Ahmed, and A. Salih, "Cloud Computing Virtualization of Resources Allocation for Distributed Systems," Journal of Applied Science and Technology Trends, vol. 1, no. 2, pp. 98–105, Jun. 2020, doi: 10.38094/jastt1331.

[4] S. Chouliaras and S. Sotiriadis, "An adaptive auto-scaling framework for cloud resource provisioning," Future Generation Computer Systems, vol. 148, pp. 173–183, 2023.

[5] I. Cohen, C. F. Chiasserini, P. Giaccone, and G. Scalosub, "Dynamic Service Provisioning in the Edge-Cloud Continuum With Bounded Resources," IEEE/ACM Transactions on Networking, vol. 31, no. 6, pp. 3096–3111, Dec. 2023, doi: 10.1109/TNET.2023.3271674.

[6] A. Abouaomar, S. Cherkaoui, Z. Mlika, and A. Kobbane, "Resource Provisioning in Edge Computing for Latency-Sensitive Applications," IEEE Internet Things J., vol. 8, no. 14, pp. 11088–11099, Jul. 2021, doi: 10.1109/JIOT.2021.3052082.

[7] D. Ardagna, M. Ciavotta, R. Lancellotti, and M. Guerriero, "A Hierarchical Receding Horizon Algorithm for QoS-Driven Control of Multi-IaaS Applications," IEEE Transactions on Cloud Computing, vol. 9, no. 2, pp. 418–434, Apr. 2021, doi: 10.1109/TCC.2018.2875443.

[8] T. Le Duc, R. G. Leiva, P. Casari, and P.-O. Östberg, "Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing," ACM Comput. Surv., vol. 52, no. 5, pp. 1–39, Sep. 2020, doi: 10.1145/3341145.

[9] K. L. James, N. P. Randall, and N. R. Haddaway, "A methodology for systematic mapping in environmental sciences," Environ. Evid., vol. 5, no. 1, p. 7, Dec. 2016, doi: 10.1186/s13750-016-0059-6.

[10] J. Zhang, H. Huang, and X. Wang, "Resource provision algorithms in cloud computing: A survey," Journal of network and computer applications, vol. 64, pp. 23–42, 2016.

[11] S. Vasoya, L. Gadhavi, J. Bhatia, and M. Bhavsar, "Resource provisioning strategies in cloud: a survey," 2016.

[12] S. Singh and I. Chana, "Cloud resource provisioning: survey, status and future research directions," Knowl. Inf. Syst., vol. 49, no. 3, pp. 1005–1069, 2016.

[13] K. D. Kumar and E. Umamaheswari, "Resource provisioning in cloud computing using prediction models: A survey," International Journal of Pure and Applied Mathematics, vol. 119, no. 9, pp. 333–342, 2018.

[14] S. Varshney, R. Sandhu, and P. K. Gupta, "QoS based resource provisioning in cloud computing environment: a technical survey," in International conference on advances in computing and data sciences, 2019, pp. 711–723.

[15] K. Sumalatha and M. S. Anbarasi, "A review on various optimization techniques of resource provisioning in cloud computing.," International Journal of Electrical & Computer Engineering (2088-8708), vol. 9, no. 1, 2019.

[16] A. Shakarami, H. Shakarami, M. Ghobaei-Arani, E. Nikougoftar, and M. Faraji-Mehmandar, "Resource provisioning in edge/fog computing: A comprehensive and systematic review," Journal of Systems Architecture, vol. 122, p. 102362, 2022.

[17] X. Li, L. Pan, and S. Liu, "A survey of resource provisioning problem in cloud brokers," Journal of Network and Computer Applications, vol. 203, p. 103384, 2022.

[18] S. Mustafa, B. Nazir, A. Hayat, S. A. Madani, and others, "Resource management in cloud computing: Taxonomy, prospects, and challenges," Computers & Electrical Engineering, vol. 47, pp. 186–203, 2015.

[19] R. S. S. Dittakavi, "An extensive exploration of techniques for resource and cost management in contemporary cloud computing environments," Applied Research in Artificial Intelligence and Cloud Computing, vol. 4, no. 1, pp. 45–61, 2021.

[20] B. H. Bhavani and H. S. Guruprasad, "Resource provisioning techniques in cloud computing environment: a survey," International Journal of Research in Computer and Communication Technology, vol. 3, no. 3, pp. 395–401, 2014.

[21] X. Xu, R. Mo, F. Dai, W. Lin, S. Wan, and W. Dou, "Dynamic resource provisioning with fault tolerance for data-intensive meteorological workflows in cloud," IEEE Trans. Industr. Inform., vol. 16, no. 9, pp. 6172–6181, 2019.

[22] K. S. S. Kumar and N. Jaisankar, "An automated resource management framework for minimizing SLA violations and negotiation in collaborative cloud," International Journal of Cognitive Computing in Engineering, vol. 1, pp. 27–35, 2020.

[23] R. Moreno-Vozmediano, R. S. Montero, E. Huedo, and I. M. Llorente, "Efficient resource provisioning for elastic cloud services based on machine learning techniques," Journal of Cloud Computing, vol. 8, no. 1, pp. 1–18, 2019.

[24] Y. Ge, Z. Ding, M. Tang, and Y.-C. Tian, "Resource provisioning for mapreduce computation in cloud container environment," in 2019 IEEE 18th international symposium on network computing and applications (NCA), 2019, pp. 1–4.

[25] S. Talwani et al., "Machine-learning-based approach for virtual machine allocation and migration," Electronics (Basel)., vol. 11, no. 19, p. 3249, 2022.

[26] M. H. Eawna, S. H. Mohammed, and E.-S. M. El-Horbaty, "Hybrid algorithm for resource provisioning of multi-tier cloud computing," Procedia Comput. Sci., vol. 65, pp. 682–690, 2015.

[27] G. Zhou, W. Tian, R. Buyya, R. Xue, and L. Song, "Deep reinforcement learning-based methods for resource scheduling in cloud computing: a review and future directions," Artif. Intell. Rev., vol. 57, no. 5, p. 124, Apr. 2024, doi: 10.1007/s10462-024-10756-9.

[28] Z. Chen, J. Hu, G. Min, C. Luo, and T. El-Ghazawi, "Adaptive and Efficient Resource Allocation in Cloud Datacenters Using Actor-Critic Deep Reinforcement Learning," IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 8, pp. 1911–1923, Aug. 2022, doi: 10.1109/TPDS.2021.3132422.

[29] D. Li, S. Xu, and P. Li, "Deep Reinforcement Learning-Empowered Resource Allocation for Mobile Edge Computing in Cellular V2X Networks," Sensors, vol. 21, no. 2, p. 372, Jan. 2021, doi: 10.3390/s21020372.

[30] W. Wei, H. Gu, K. Wang, J. Li, X. Zhang, and N. Wang, "Multi-Dimensional Resource Allocation in Distributed Data Centers Using Deep Reinforcement Learning," IEEE Transactions on Network and Service Management, vol. 20, no. 2, pp. 1817–1829, Jun. 2023, doi: 10.1109/TNSM.2022.3213575.

[31] A. Thomas, T. Borangiu, and D. Trentesaux, "Holonic and multi-agent technologies for service and computing oriented manufacturing," J. Intell. Manuf., vol. 28, no. 7, pp. 1501–1502, Oct. 2017, doi: 10.1007/s10845-015-1188-4.

[32] B. Fang and D. Gao, "Collaborative Multi-Agent Reinforcement Learning Approach for Elastic Cloud Resource Scaling," in 2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA), IEEE, Jun. 2025, pp. 415–419. doi: 10.1109/ICAITA67588.2025.11137847.

[33] N. Šatkauskas and A. Venčkauskas, "Multi-Agent Dynamic Fog Service Placement Approach," Future Internet, vol. 16, no. 7, p. 248, Jul. 2024, doi: 10.3390/fi16070248.

[34] P. L. Donti, B. Amos, and J. Z. Kolter, "Task-based End-to-end Model Learning in Stochastic Optimization."

[35] J. Liu, Y. Zhang, Y. Zhou, D. Zhang, and H. Liu, "Aggressive Resource Provisioning for Ensuring QoS in Virtualized Environments," IEEE Transactions on Cloud Computing, vol. 3, no. 2, pp. 119–131, Apr. 2015, doi: 10.1109/TCC.2014.2353045.

[36] Z. Xiao, W. Song, and Q. Chen, "Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment," IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 6, pp. 1107–1117, Jun. 2013, doi: 10.1109/TPDS.2012.283.

[37] Prof. P. R. P. Prof. P. R. Patil, M. R. Manasi Raut, A. B. Amruta Bakade, and A. S. Anushka Shingade, "Analysis of Virtual Machine Allocation Strategies and Development of a Novel Policy," Journal of Software Engineering and Simulation, vol. 11, no. 5, pp. 01–14, May 2025, doi: 10.35629/3795-11050114.

[38] S. Liu, C. Li, Z. Liu, and Q. Zhang, "Virtual Machine Dynamic Deployment Scheme Based on Double-Cursor Mechanism," IEEE Access, vol. 8, pp. 214481–214493, 2020, doi: 10.1109/ACCESS.2020.3040912.

[39] A. Laghrissi and T. Taleb, "A Survey on the Placement of Virtual Resources and Virtual Network Functions," IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1409–1434, 2019, doi: 10.1109/COMST.2018.2884835.

[40] W. Wei, K. Wang, K. Wang, H. Gu, and H. Shen, "Multi-resource balance optimization for virtual machine placement in cloud data centers," Computers & Electrical Engineering, vol. 88, p. 106866, Dec. 2020, doi: 10.1016/j.compeleceng.2020.106866.

[41] N. Zhou, H. Zhou, and D. Hoppe, "Containerization for High Performance Computing Systems: Survey and Prospects," IEEE Transactions on Software Engineering, vol. 49, no. 4, pp. 2722–2740, Apr. 2023, doi: 10.1109/TSE.2022.3229221.

[42] M. A. Rodriguez and R. Buyya, "Container‐based cluster orchestration systems: A taxonomy and future directions," Softw. Pract. Exp., vol. 49, no. 5, pp. 698‐719, May 2019, doi: 10.1002/spe.2660.

[43] Z. Zhong, M. Xu, M. A. Rodriguez, C. Xu, and R. Buyya, "Machine Learning-based Orchestration of Containers: A Taxonomy and Future Directions," ACM Comput. Surv., vol. 54, no. 10s, pp. 1–35, Jan. 2022, doi: 10.1145/3510415.

[44] Z. Zhong and R. Buyya, "A Cost-Efficient Container Orchestration Strategy in Kubernetes-Based Cloud Computing Infrastructures with Heterogeneous Resources," ACM Trans. Internet Technol., vol. 20, no. 2, pp. 1–24, May 2020, doi: 10.1145/3378447.

[45] M. A. Altahat, T. Daradkeh, and A. Agarwal, "Optimized Encryption-Integrated Strategy for Containers Scheduling and Secure Migration in Multi-Cloud Data Centers," IEEE Access, vol. 12, pp. 51330–51345, 2024, doi: 10.1109/ACCESS.2024.3386169.

[46] V. Struhár, S. S. Craciunas, M. Ashjaei, M. Behnam, and A. V. Papadopoulos, "Hierarchical Resource Orchestration Framework for Real-time Containers," ACM Transactions on Embedded Computing Systems, vol. 23, no. 1, pp. 1–24, Jan. 2024, doi: 10.1145/3592856.

[47] Z. Li et al., "Online Layer-Aware Joint Request Scheduling, Container Placement, and Resource Provision in Edge Computing," IEEE Trans. Serv. Comput., vol. 18, no. 1, pp. 328–341, Jan. 2025, doi: 10.1109/TSC.2024.3504237.

[48] Grace Joseph, Sunandha Rajagopal, Dr. Amrita Priya K, and Sreelekshmi R, "Dynamic Resource Scheduling Approaches in Server Less Computing," International Research Journal on Advanced Engineering and Management (IRJAEM), vol. 3, no. 05, pp. 1749–1758, May 2025, doi: 10.47392/IRJAEM.2025.0277.

[49] A. Mampage, S. Karunasekera, and R. Buyya, "A Holistic View on Resource Management in Serverless Computing Environments: Taxonomy and Future Directions," ACM Comput. Surv., vol. 54, no. 11s, pp. 1–36, Jan. 2022, doi: 10.1145/3510412.

[50] A. Mampage, S. Karunasekera, and R. Buyya, "Deadline-aware Dynamic Resource Management in Serverless Computing Environments," in 2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid), IEEE, May 2021, pp. 483–492. doi: 10.1109/CCGrid51090.2021.00058.

[51] P. Benedetti, M. Femminella, G. Reali, and K. Steenhaut, "Experimental Analysis of the Application of Serverless Computing to IoT Platforms," Sensors, vol. 21, no. 3, p. 928, Jan. 2021, doi: 10.3390/s21030928.

[52] H. Yu, H. Wang, J. Li, X. Yuan, and S.-J. Park, "Accelerating Serverless Computing by Harvesting Idle Resources," in Proceedings of the ACM Web Conference 2022, New York, NY, USA: ACM, Apr. 2022, pp. 1741–1751. doi: 10.1145/3485447.3511979.

[53] B. Zhu, Y. Zhu, C. Chen, and L. Kong, "Trident: A Provider-Oriented Resource Management Framework for Serverless Computing Platforms," IEEE Trans. Serv. Comput., vol. 18, no. 5, pp. 3334–3347, Sep. 2025, doi: 10.1109/TSC.2025.3603867.

[54] S. Rac and M. Brorsson, "Cost-aware Service Placement and Scheduling in the Edge-Cloud Continuum," ACM Transactions on Architecture and Code Optimization, vol. 21, no. 2, pp. 1–24, Jun. 2024, doi: 10.1145/3640823.

[55] I. Capeletti et al., "Towards Optimizing the Edge-to-Cloud Continuum Resource Allocation," in Proceedings of the 13th International Conference on Cloud Computing and Services Science, SCITEPRESS - Science and Technology Publications, 2023, pp. 90–99. doi: 10.5220/0011995700003488.

[56] A. Bertoncini, A. Ceselli, and C. Quadri, "Latency-Aware Placement of Microservices in the Cloud-to-Edge Continuum via Resource Scaling," in 2025 IEEE International Conference on Smart Computing (SMARTCOMP), IEEE, Jun. 2025, pp. 420–425. doi: 10.1109/SMARTCOMP65954.2025.00103.

[57] M. I. Hussain et al., "Hybrid SFLA-UBS algorithm for optimal resource provisioning with cost management in multi-cloud computing," International Journal of Advanced Computer Science and Applications, vol. 12, no. 4, 2021.

[58] M. Kumar, A. Kishor, J. Abawajy, P. Agarwal, A. Singh, and A. Y. Zomaya, "ARPS: An autonomic resource provisioning and scheduling framework for cloud platforms," IEEE Transactions on Sustainable Computing, vol. 7, no. 2, pp. 386–399, 2021.

[59] M. Ghobaei-Arani and A. Shahidinejad, "An efficient resource provisioning approach for analyzing cloud workloads: a metaheuristic-based clustering approach," J. Supercomput., vol. 77, no. 1, pp. 711–750, 2021.

[60] A. Suresh and R. Varatharajan, "Competent resource provisioning and distribution techniques for cloud computing environment," Cluster Comput., vol. 22, no. Suppl 5, pp. 11039–11046, 2019.

[61] M. M. Nejad, L. Mashayekhy, and D. Grosu, "Truthful greedy mechanisms for dynamic virtual machine provisioning and allocation in clouds," IEEE transactions on parallel and distributed systems, vol. 26, no. 2, pp. 594–603, 2014.

[62] R. Panwar and M. Supriya, "Dynamic resource provisioning for service-based cloud applications: A Bayesian learning approach," J. Parallel Distrib. Comput., vol. 168, pp. 90–107, 2022.

[63] C. Luo et al., "Intelligent virtual machine provisioning in cloud computing," in Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 1495–1502.

[64] S. A. Ajila and A. A. Bankole, "Using machine learning algorithms for cloud client prediction models in a web VM resource provisioning environment," Transactions on Machine Learning and Artificial Intelligence, vol. 4, no. 1, p. 28, 2016.

[65] F. Mirzad* and M. R. Ghalib, "Forecasting Cloud Resource Provisioning System using Supervised Machine Learning," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 6, pp. 3591–3596, Mar. 2020, doi: 10.35940/ijrte.F8886.038620.

[66] Y. Sharma, J. Taheri, W. Si, D. Sun, and B. Javadi, "Dynamic resource provisioning for sustainable cloud computing systems in the presence of correlated failures," IEEE Transactions on Sustainable Computing, vol. 6, no. 4, pp. 641–654, 2020.

[67] B. Sniezynski, P. Nawrocki, M. Wilk, M. Jarzab, and K. Zielinski, "VM reservation plan adaptation using machine learning in cloud computing," J. Grid Comput., vol. 17, no. 4, pp. 797–812, 2019.

[68] M. S. Al-Asaly, M. A. Bencherif, A. Alsanad, and M. M. Hassan, "A deep learning-based resource usage prediction model for resource provisioning in an autonomic cloud computing environment," Neural Comput. Appl., vol. 34, no. 13, pp. 10211–10228, 2022.

[69] J. Choi and Y. Kim, "Adaptive resource provisioning method using application-aware machine learning based on job history in heterogeneous infrastructures," Cluster Comput., vol. 20, no. 4, pp. 3537–3549, 2017.

[70] S. Ghasemi, M. R. Meybodi, M. D. T. Fooladi, and A. M. Rahmani, "A cost-aware mechanism for optimized resource provisioning in cloud computing," Cluster Comput., vol. 21, no. 2, pp. 1381–1394, 2018.

[71] M. P. Yadav, Rohit, and D. K. Yadav, "Resource provisioning through machine learning in cloud services," Arab. J. Sci. Eng., vol. 47, no. 2, pp. 1483–1505, 2022.

[72] C. Ayimba, P. Casari, and V. Mancuso, "SQLR: Short-term memory Q-learning for elastic provisioning," IEEE Transactions on Network and Service Management, vol. 18, no. 2, pp. 1850–1869, 2021.

[73] S. Agarwal, M. A. Rodriguez, and R. Buyya, "A Deep Recurrent-Reinforcement Learning Method for Intelligent AutoScaling of Serverless Functions," IEEE Trans. Serv. Comput., vol. 17, no. 5, pp. 1899–1910, 2024, doi: 10.1109/TSC.2024.3387661.

[74] H. Yu, H. Wang, J. Li, X. Yuan, and S.-J. Park, "Freyr +: Harvesting Idle Resources in Serverless Computing via Deep Reinforcement Learning," IEEE Transactions on Parallel and Distributed Systems, vol. 35, no. 11, pp. 2254–2269, Nov. 2024, doi: 10.1109/TPDS.2024.3462294.

[75] H. Bai, M. Xu, K. Ye, R. Buyya, and C. Xu, "DRPC: Distributed Reinforcement Learning Approach for Scalable Resource Provisioning in Container-Based Clusters," IEEE Trans. Serv. Comput., vol. 17, no. 6, pp. 3473–3484, 2024, doi: 10.1109/TSC.2024.3433388.

[76] S. Nagarajan, P. S. Rani, M. S. Vinmathi, V. Subba Reddy, A. L. M. Saleth, and D. Abdus Subhahan, "Multi agent deep reinforcement learning for resource allocation in container-based clouds environments," Expert Syst., vol. 42, no. 1, Jan. 2025, doi: 10.1111/exsy.13362.

[77] T. Theodoropoulos, A. Makris, I. Korontanis, and K. Tserpes, "GreenKube: Towards Greener Container Orchestration using Artificial Intelligence." [Online]. Available: https://kubernetes.io/

[78] M. Golec et al., "ATOM: AI-Powered Sustainable Resource Management for Serverless Edge Computing Environments," IEEE Transactions on Sustainable Computing, vol. 9, no. 6, pp. 817–829, 2024, doi: 10.1109/TSUSC.2023.3348157.

[79] H. Zhang, J. Wang, H. Zhang, and C. Bu, "Security computing resource allocation based on deep reinforcement learning in serverless multi-cloud edge computing," Future Generation Computer Systems, vol. 151, pp. 152–161, Feb. 2024, doi: 10.1016/j.future.2023.09.016.

[80] X. Chen, L. Yang, Z. Chen, G. Min, X. Zheng, and C. Rong, "Resource Allocation With Workload-Time Windows for Cloud-Based Software Services: A Deep Reinforcement Learning Approach," IEEE Transactions on Cloud Computing, vol. 11, no. 2, pp. 1871–1885, Apr. 2023, doi: 10.1109/TCC.2022.3169157.

[81] N. Filinis et al., "Intent-driven orchestration of serverless applications in the computing continuum," Future Generation Computer Systems, vol. 154, pp. 72–86, May 2024, doi: 10.1016/j.future.2023.12.032.

[82] Y. Gong, J. Huang, B. Liu, J. Xu, B. Wu, and Y. Zhang, "Dynamic Resource Allocation for Virtual Machine Migration Optimization using Machine Learning," Applied and Computational Engineering, vol. 57, no. 1, pp. 1–8, Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.13619

[83] Y. Wang and X. Yang, "Intelligent Resource Allocation Optimization for Cloud Computing via Machine Learning," Advances in Computer, Signals and Systems, vol. 9, no. 1, Mar. 2025, doi: 10.23977/acss.2025.090109.

[84] A. Lipari, G. P. Mattia, and R. Beraldi, "Dynamic and Forecast-Based Containers Autoscaling for Kubernetes with Reinforcement Learning," in Proceedings - 2025 IEEE International Parallel and Distributed Processing Symposium Workshops, IPDPSW 2025, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 1081–1088. doi: 10.1109/IPDPSW66978.2025.00169.

[85] A. Mampage, S. Karunasekera, and R. Buyya, "Deep Reinforcement Learning for Scheduling Applications in Serverless and Serverful Hybrid Computing Environments," IEEE Trans. Serv. Comput., vol. 18, no. 2, pp. 718–728, 2025, doi: 10.1109/TSC.2024.3520864.

[86] P. Wu, H. Chen, T. Wu, K. Gu, and Y. Xia, "Cold-Start-Aware Offloading and Resource Allocation by Importance Sampling-Based Double Dueling DQN in Serverless Edge Computing," IEEE Internet Things J., vol. 12, no. 19, pp. 40190–40205, 2025, doi: 10.1109/JIOT.2025.3588727.

A Machine Learning-Driven Framework for Accurate Brain Image Registration in Multimodal and Noisy Environments

M. S. Minu[1], Mutharasu M[2], S. Hemamalini[3], Sunitha T[4], Mohanaprakash T A[5], Justindhas Y[6]
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India[1]
Assistant Professor, Department of CSE (Cyber Security), Madanapalle Institute of Technology & Science, Andhra Pradesh, India[2]
Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India, Tamil Nadu - 600 062[3]
Assistant Professor, Department of Artificial Intelligence and Data Science, Saveetha Engineering College, Thandalam, India[4]
Department of CSE, R. M. K. Engineering College, Kavaraipettai, Chennai, India[5]
Assistant Professor, Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Easwari Engineering College, Ramapuram, Chennai, India[6]

***Abstract*—Brain image registration is fundamental for medical imaging to allow the matching of images from multiple modalities, temporal sequences, and people to offer spatial correlation. This is crucial for activities such as cohort studies, intervention recommendations, and treatment monitoring, where exact alignment assures consistent analysis. Notwithstanding their importance, modern brain image registration techniques have many shortcomings, including limited resistance to noise, misalignment in multi-modality images, and costly computational expenses. These limits may impede real-time clinical environment practical implementation and provide less than optimal registration accuracy. This study addresses these issues by means of an Improved Brain Image Registration Technique Using Machine Learning Algorithms (BIRT-MLA). The proposed architecture detects significant image properties by means of convolutional neural networks (CNNs), therefore enabling feature extraction. By applying a supervised learning method, it guarantees precise alignment even in noisy and demanding imaging situations by forecasting transformation parameters. Lowering the registration error by modern optimization techniques helps to save processing time and maintain remarkable accuracy even in this respect. Using CNNs, the proposed method helps to effectively classify brain images, thereby improving diagnostic support and the usefulness of registered images for downstream operations. Improving clinical judgment and simplifying processes rely on grouping registration and categorization into a logical sequence. By means of enhanced alignment precision, resistance to picture faults, and shortened computing time compared to current approaches, experimental findings expose the advantages of the suggested technology. This development may be very useful in clinical and experimental settings, thereby supporting the accuracy and efficiency of brain picture analysis.**

Keywords*—Machine learning algorithms; brain image registration; convolutional neural networks; deep learning; brain disease

I. Introduction

A. Background

Medical image registration, specifically that of the brain, has become an essential aspect of modern medical imaging [1]. It enables images to be aligned across various modalities, time points, and subjects. For accurate picture analysis in various medical practices, especially in the multi-modal medical aids, the spatial alignment of overlapping material is always a prerequisite [2]. Brain imaging in cohort studies, treatment-guided studies or therapy evaluative progress has one way or the other tried to rely on these brain imaging registrations [3]. To make the most helpful conclusions, assess differences, and adjust methods over time, other activities must be performed with great detail and consistency, especially the acquisition of the images [4]. Considering the increasing volume of multimodal brain imaging data together with the recent advances in imaging technologies, the requirement for robust and efficient image registration techniques has emerged [5]. However, current techniques cannot avoid noise, misalignment, and the complexity of multiple images [6]. Furthermore, the conventional techniques introduce traces of overhead which are a challenge to work with in a clinical environment owing to the high demand for image processing that is both accurate and quick [7].

Integrating machine learning into brain image registration provides promising answers to the numerous questions raised by this area of research [8]. Most especially convolutional neural networks but also other ML methods excel at feature extraction and even pattern recognition which is suitable for the challenging imaging data [9]. This has been made possible by ML which facilitates an improvement in the accuracy, robustness and efficiency of the brain image registration technique [10].

This study presents an Improved Brain Image Registration Technique Using Machine Learning Algorithms (BIRT-MLA), which is intended to improve on the drawbacks of the existing methods and also contribute to the development of approaches to brain imaging [11]. As the reliance on multimodal imaging in clinical and research settings continues to grow, improving the efficiency and accuracy of this registration process becomes even more critical [12]. When applied to real-time scenarios, traditional methods tend to be hampered by noise, high computing cost, and misalignment, and these are the same problems they encounter [13].

Motivation: Machine learning provides one opportunity of dealing with these issues, hence enabling building of high and robust systems for brain image registration, which in turn improves the diagnosis and treatment monitoring [14].

Problem statement: Many of the contemporary brain imaging transformations' solutions lack stability, accuracy, and computational efficiency, which makes their application difficult in clinical practice [15]. Such challenges encompass the sensitivity to noise, registration of multi-modality images, and prolonged response times [16]. These constraints make them less practically useful for real-time analysis and formulation of feedback, as well as give not the best registration results [17]. Enhancement of wider applicability and enhanced efficiency of brain image registration has to do with these problems.

B. Goal of this Paper

The evolution of the BIRT-MLA framework has inspired the idea of a new method based on machine learning for brain image registration. Using convolutional neural networks helps this system to provide strong feature extraction and perfect alignment.

By means of their decrease in the number of errors during the registration process as well as the expenditure of money on processing, the advanced optimization solutions discussed here serve to expand the applicability of the system in real-time environments.

C. Problem Statement and Significance

The simplification of clinical processes and the rise of diagnostic capacities are much influenced by the mix of registration and classification approaches.

The remainder of this study is structured as follows: In Section II, the related work of brain image registration is studied. In Section III, the proposed methodology of BIRT-MLA is explained. In Section IV, the efficiency of BIRT-MLA is discussed and analyzed. Finally, in Section V, the study is concluded with the future work in Section VI.

II. Related Work

Analyzing deep learning-based medical image registration approaches with an eye on advancements, obstacles, and new trends, it covers organ atlas creation and tumor growth monitoring, as well as the progress of DL-MIR and classifies techniques with reference to their therapeutic usefulness. Apart from interesting findings, the study shows future opportunities to improve clinical relevance and efficiency.

A. Comprehensive DL-MIR Review

Emphasizing those with significant methodological or functional breakthroughs, the present analysis looks at all deep learning-based medical image registration papers released since the present. It analyzes the use of deep learning methods in medical photo registration, therefore summarizing growing trends and problems [18]. The study identifies unmet clinical needs and offers ideas on potential future ways to advance the subject, thus ensuring relevance to real-world clinical applications and enabling the continuous advancement of image registration techniques.

B. DL-MIR Evolution Survey

The study methodically investigates the evolution of deep learning-based medical image registration across many years, including assessment of developments, challenges, and research patterns. Emphasizing relevant research challenges faced by practitioners, it focuses on clinical activities such as tumor growth monitoring, image fusion, and organ atlas development [19]. This study summarizes significant advancements, clarifies the quick adoption of deep learning, and offers future directions to meet unmet needs and forward the field of medical image registration to new degrees of usefulness.

C. DL-MIR Review with Categorization

In the present analysis, an overview of deep learning (DL)-based medical image registration methods is offered. We covered the most current developments in the medical sector of DL-based registration systems. Their methods, goals, and popularity led to dividing these strategies into seven categories [20]. Every area was carefully examined to highlight specific challenges and reveal notable accomplishments. Following the exhaustive study of every subject, a quick assessment was provided to underline its achievements and future prospects. We provided a complete comparison of DL-based techniques for brain and lung registration using benchmark data. At last, we investigated the data of every indicated research from many viewpoints to reveal the popularity and future direction of growth of DL-based medical image registration.

D. ML-DL Brain Disease Review

This effort compiles 147 recent works on machine learning and deep learning approaches for Parkinson's illness, brain tumors, epilepsy, and Alzheimer's disease detection. Analyzing 22 often-used datasets and feature extraction techniques in diagnosing brain diseases, it evaluates several methodologies, data modalities, and datasets [21]. The study discusses significant findings, points out challenges, and investigates the potential of artificial intelligence-based diagnostics. The objective is to highlight the most accurate techniques for the diagnosis of brain illnesses and provide recommendations for further advances in the field.

E. Deep Learning-Based Medical Image Registration

Under supervised, unsupervised, and iterative frameworks, deep learning-based medical image registration methods are investigated in this study. Among supervised methods are dual, completely, and weakly supervised registration [22].

Unsupervised methods rely on similarity-based and GAN-based registration techniques. Deep iterative registration brings deep similarity learning along with reinforcement learning. Stressing the lack of training datasets with known transformations as a main challenge, the work also explores monomodal and multimodal image registration across modalities, including X-ray, CT, ultrasonic, and MRI, therefore directing future research directions.

F. Deep Neural Networks for Medical Image Registration

This work rigorously assesses medical image registration using deep neural networks (MIR-DNN), underlining improvements enabled by many-core CPUs. Emphasizing the transformational potential of deep learning, it looks at modern methods, basic concepts, statistical analysis, and challenges [23]. Providing researchers with all the information they need, the review addresses original contributions, enabling strategies, and possible advances. The major focus is on the improvement of registration efficiency and effectiveness for image-guided treatments, thereby providing a foundation for further field research and development.

G. DL-Based Lung Image Registration

Aiming at DL-based lung image registration (DL-LIR), this work resolves problems like major deformities and motion discontinuities resulting from breathing and heart activity. It ranks techniques in three fully-supervised, weakly-supervised, and unsupervised categories [24]. It undertakes a comprehensive statistical analysis using publicly available datasets, loss functions, and measures. By means of a combination of conventional and DL-based techniques, the review highlights unresolved problems and proposes future paths to improve lung image registration accuracy and usefulness in clinical environments.

H. Learn2Reg Framework

The Learn2Reg challenge offers a consistent framework for evaluation of deformable 3D medical image registration systems across several challenges, anatomies (brain, abdomen, thorax), and imaging modalities (CT, MR, ultrasonic) as (L2R) Combining complementary measurements such as accuracy, resilience, plausible behavior, and runtime lets one assess registration performance all around. Emphasizing multi-task training, label monitoring, and transferability analysis, the challenge helps one to understand the trade-offs of conventional and deep-learning-based approaches [25]. This method helps experts evaluate and improve contemporary registration methods on therapeutically relevant data.

By means of thorough analysis of DL-MIR techniques, the research divides them into iterative, unsupervised, and supervised models. It underlines issues such as data limitation and distortion in lung imaging. Analyzing significant studies in many different disciplines allows the study to identify trends, unmet needs, and provide ideas for future advances to enhance therapeutic value and accuracy in medical image registration. Table I presents a comparison of existing methods.

TABLE I. THE COMPARISON OF EXISTING METHODS

S. No	Methods	Advantages	Limitations
1	Comprehensive DL-MIR Review	Summarizes growing trends and challenges in DL-MIR, with a focus on real-world clinical applications	Identifies unmet clinical needs but lacks specific solutions for certain challenges
2	DL-MIR Evolution Survey	Highlights significant advancements, quick adoption of deep learning, and offers future directions	Focuses on research challenges but may not offer practical solutions for all clinical issues
3	DL-MIR Review with Categorization	Provides an extensive overview and categorization of DL-based methods, comparing brain and lung registration	The broad categorization may not cover niche or emerging methods comprehensively
4	ML-DL Brain Disease Review	Provides a comprehensive compilation of machine learning and deep learning methods for brain disease detection	Challenges in standardizing datasets and the variety of methods lead to inconsistent results
5	Deep Learning-Based Medical Image Registration	Explores supervised, unsupervised, and iterative methods, focusing on multimodal image registration	Lack of training datasets with known transformations and challenges with multimodal registration
6	Deep Neural Networks for Medical Image Registration	Assesses the improvement enabled by deep learning, especially in image-guided treatments	Efficiency issues in complex multi-core CPU environments may limit practical application
7	DL-Based Lung Image Registration	Addresses issues of deformities and motion discontinuities in lung image registration	Motion artifacts and issues with breathing and heart activity continue to challenge accuracy
8	Learn2Reg Framework	Offers a consistent evaluation framework for 3D deformable image registration across challenges and modalities	Complex multi-task training and transferability analysis may be difficult to implement in clinical environments

III. PROPOSED METHODOLOGY

This work illustrates how advanced machine learning techniques might be added to brain image registration. Investigating many deep learning approaches, including CNNs, GANs, and U-Net, helps to increase the accuracy and efficiency of medical image alignment, thus allowing better comparisons for clinical and research purposes. By means of low error and processing time, the system aims to improve real-time clinical applications.

A. Contribution 1: Proposed a Novel Machine Learning-Based Registration Framework

Presented a strong brain image registration method using convolutional neural networks (CNNs) for accurate feature extraction and supervised learning to estimate transformation parameters, therefore guaranteeing exact alignment even in noisy and challenging imaging environments.

1) Learning approach: Fig. 1 explains the brain image registration process of advanced machine learning methods.

Starting with raw brain images (MRI, CT, PET Scans), the method consists of denoising and normalizing preprocessing steps. Among other conventional and deep learning methods, feature extraction takes use of cnns and autoencoders. GANs and U-Net are among the deep learning-based approaches; traditional techniques are used in registration. Following efficient alignment ensured by optimization techniques comes with model evaluation using DSC and MSE. The outcome is a registered brain image generated with high-quality fusion and statistical analysis guaranteed by post-processing methods.

$$||ex - sm''||: \rightarrow \varphi_v S[\delta + uy''] - Cqv^e[x - zk''] + Mn[c - xz''] \quad (1)$$

To improve alignment and feature correlation, Eq. (1) 1n represents the interaction of optimization terms ($||ex - sm''||$, $\varphi_v S[\delta + uy'']$ with transform spatial parameters $Cqv^e[x - zk'']$, $Mn[c - xz'']$. This aligns with the study's objective of obtaining computationally efficient, accurate, and noise-resistant brain image registration [see Eq. (2)]:

$$\varepsilon_f D[vc'':sz]: \rightarrow \in \nabla x' + 9u[\alpha\exists' - 3vap] - Ca[\varepsilon\beta - 9avs''] \quad (2)$$

Under noise and alignment adjustments, the algorithm depicts $\in \nabla x' + 9u$ the dynamic interaction of the spatial gradients ($\varepsilon_f D$) and feature constraints ($[vc'':sz]$). The suggested BIRT-MLA approach incorporates convolutional neural networks (CNNs) to improve registration accuracy and noise resistance by altering transformations ($9u[\alpha\exists' - 3vap]$) and obtaining robust features ($Ca[\varepsilon\beta - 9avs'']$. In difficult medical imaging situations, this helps the framework achieve its aim of accurate alignment and computing efficiency.

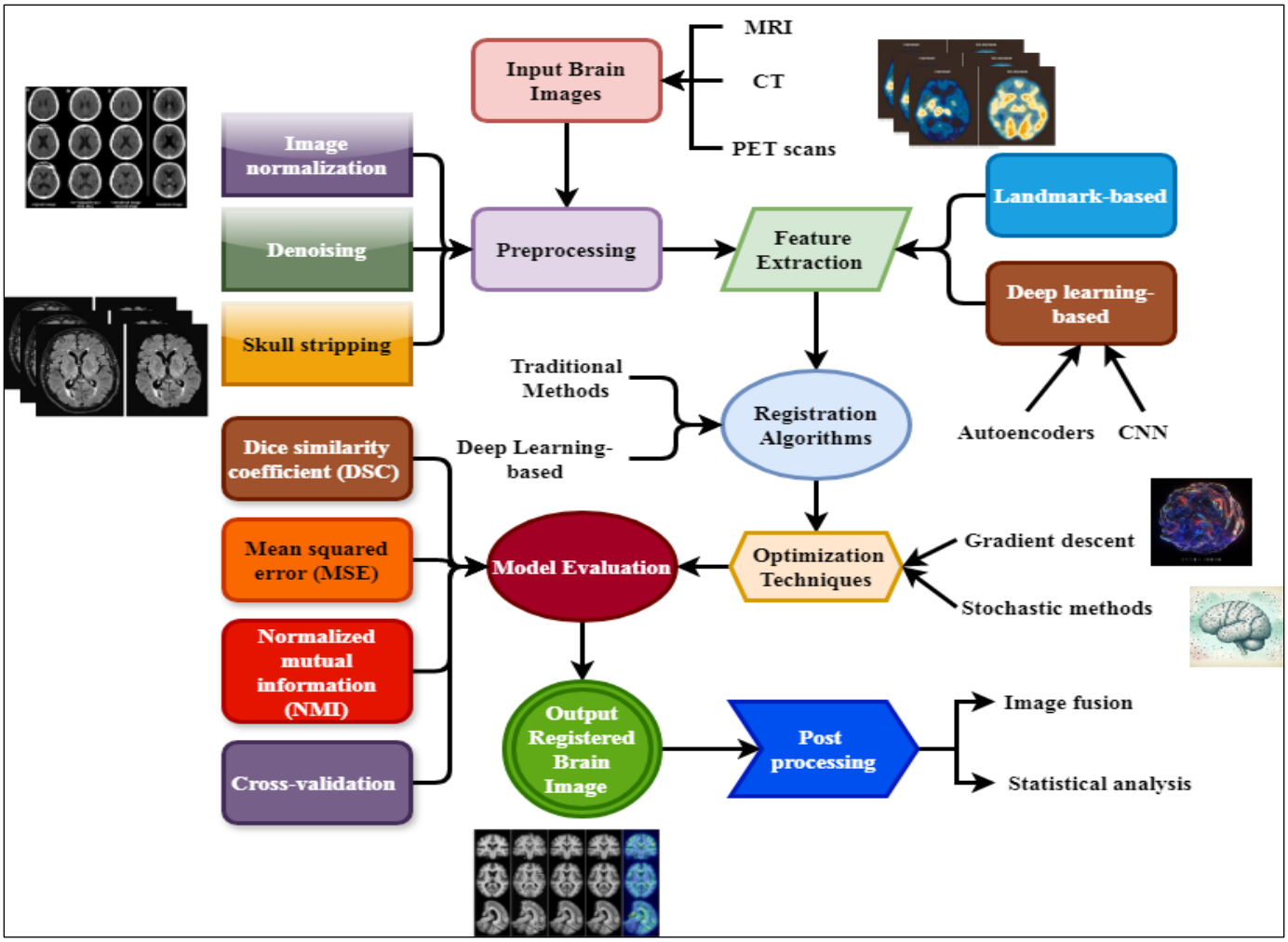


Fig. 1. AI-enhanced brain image registration: a machine.

Fig. 2 illustrates the two-step medical image registration process aligning a floating image to a reference image for perfect comparison. Linear Registration starts with global transformations including translation, rotation, scaling, shearing to achieve approximate alignment. Then polishes the result by using a deformation field to modify local variances, therefore assuring complete anatomical structural alignment. From coarse to fine-grained corrections, the method evolves to provide very accurate final pictures. Medical imaging relies on this mechanism for processes like brain analysis, which enables very accurate comparisons between patients or time points.

$$C_x S[\forall - 9bv'']: \rightarrow Ku[\forall s'' + 9vaw''] - Cva[\propto + 9suy''] \quad (3)$$

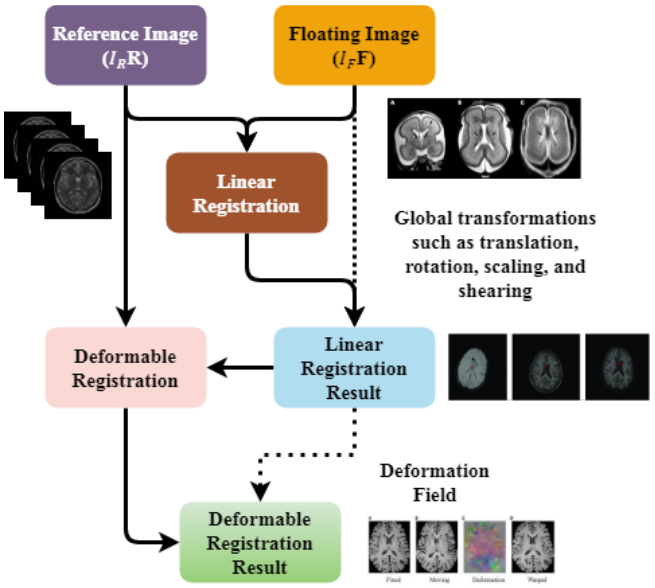


Fig. 2. Global alignment to precise deformation: image registration workflow.

Aligning multi-modal data $\forall - 9bv''$ while considering distortions $Ku[\forall s'' + 9vaw'']$ is included in Eq. (3), in which $Cva[\propto +9suy'']$ also optimizes spatial characteristics (C_xS). The goal of the approach is to improve the accuracy and usefulness of picture registration in clinical and industrial workflows, and this procedure is in line with that purpose.

$$c_{dF}[n - bx'']: \rightarrow Lk[\forall \propto' - sn] + 7t[\forall - 3en''] - Cg[\partial \forall - cz''] \quad (4)$$

To improve alignment, the coefficients of learning (c_{dF}), feature transformations ($[n - bx'']: \rightarrow$), and noise adjustments ($Cg[\partial \forall - cz'']$) interact with each other. To achieve accurate registration in the BIRT-MLA the structure convolutional neural networks (CNNs) use noise resilience $Cg[\partial \forall - cz'']$ and minimization ($[\forall \propto' - sn]$) to extract and forecast these feature-based modifications. The method's focus is on obtaining efficient and robust alignment in complex medical imaging environments for Eq. (4).

B. Contribution 2: Enhanced Efficiency and Accuracy in Brain Image Registration

The system is appropriate for real-time clinical uses as included sophisticated optimization algorithms minimize registration mistakes, greatly reduce processing time, and enhance resilience against imaging distortions. Input brain pictures from the beginning of the process and pre-processing procedures including noise reduction and normalization is shown in Fig. 3(a).

Fig. 3(b) describes a pairwise direct linear registration method for standard template alignment of a brain picture from a test participant. First aligned with many intermediary mediator pictures using affine transformations (TS→mi) the test image is Predefined a priori transformations (Tmi→t) then transfer each mediator to the ultimate template picture. Leveraging mediators as intermediary references guarantees strong and exact registration via this multi-step alignment. Forming a fundamental basis for sophisticated medical imaging technologies and machine learning applications, the framework enables correct comparison and analysis of brain anatomy across people and research [see Eq. (5)].

$$X_sF[ju - 4va'']: \rightarrow Las[\cup -5ca''] + 8h[\forall w - ab''] - Cq[bn''] \quad (5)$$

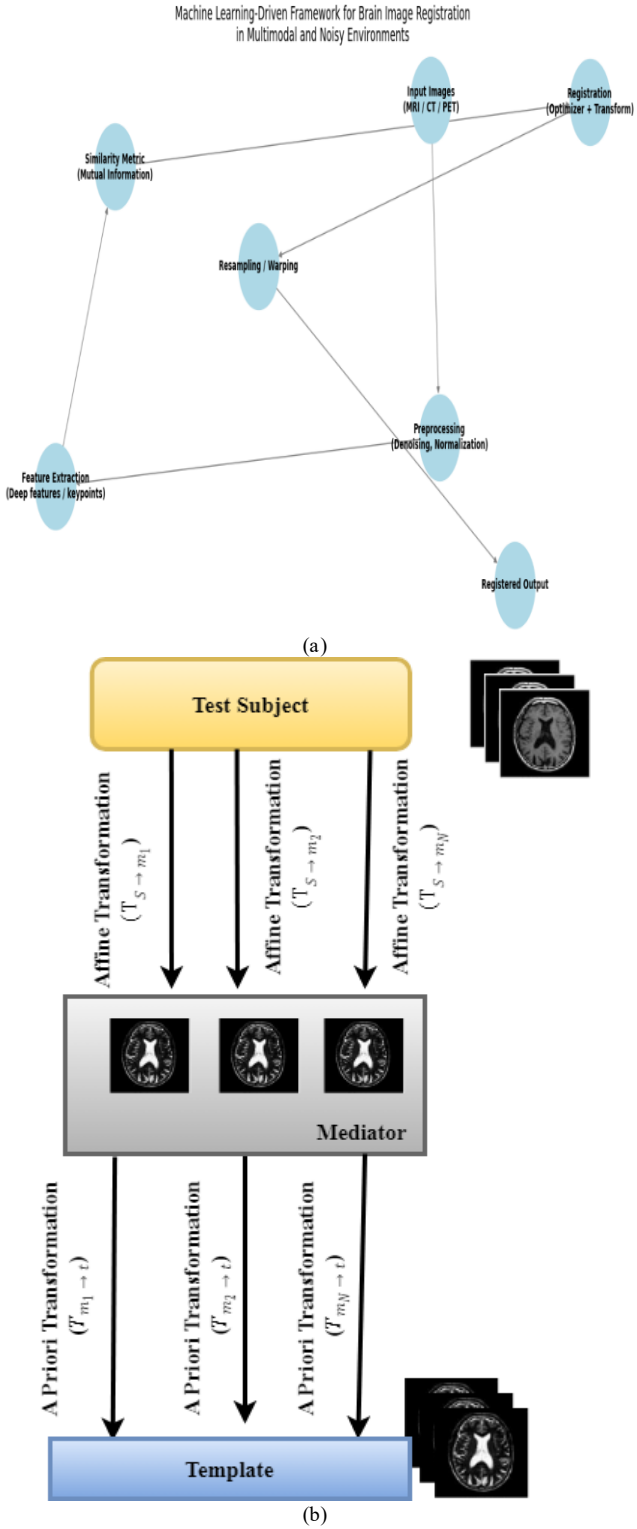


Fig. 3. (a) Procedure for dataset input image pre-processing, (b) Bridging transformations: mediator-based image alignment framework.

The optimization terms (X_sF) control the transformation $8h[\forall w - ab''] -$ of spatial characteristics ($ju - 4va''$) to compensate $Cq[bn'']$ for distortions ($Las[\cup -5ca'']$).This lends credence to the method's stated purpose of obtaining clinically and experimentally applicable brain image registration that is both dependable and robust to noise.

$$d_cD[6v - am'']: \rightarrow Ju[\forall \propto -6va''] + 7\forall[\partial \times \propto \times z''] - 9u[\forall s''] \quad (6)$$

To improve spatial alignment $Ju[\forall \propto -6va'']$, optimizing terms (d_cD) and feature transformations ($[6v - am'']$ are represented by Eq. (6). To be able to forecast accurate alignments $7\forall[\partial \times \propto \times z'']$ and guarantee strong registration, CNNs in the BIRT-MLA approach use such modifications and noise-adjusted values $9u[\forall s'']$. The method's goal of increasing precision and computing efficiency under difficult imaging circumstances is well-aligned with this.

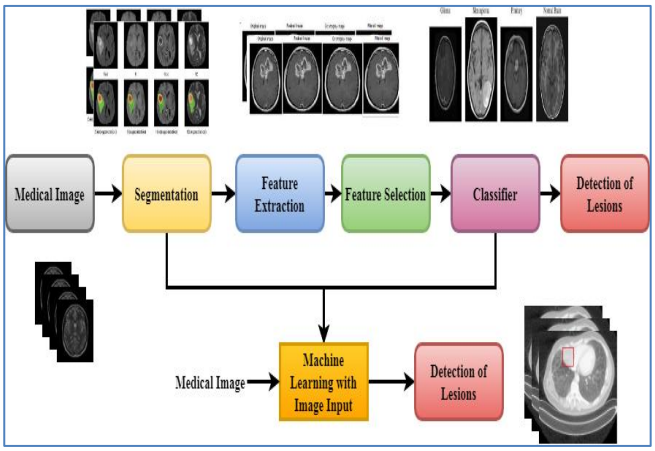


Fig. 4. From features to networks: evolution of lesion detection techniques.

Fig. 4 compares medical picture analysis classical machine learning techniques with deep learning methods. The conventional route calls for segmentation, feature extraction, feature selection, and classification to find lesions, therefore needing considerable human involvement and domain knowledge. By immediately processing medical pictures with complex neural networks such CNNs, MTANNs, and DBNs, the deep learning approach avoids these middle processes. Using data-driven learning, this simplified technique improves accuracy and efficiency. The contrast emphasizes how transforming deep learning is in automating and enhancing lesion diagnosis, hence opening the path for sophisticated medical imaging uses in clinical practice and research.

$$c_Ds[V - xv'']: \rightarrow Ku'[\partial\forall \times dv''] + Uy[x - zk''] - [\partial\forall - 2wm''] \quad (7)$$

To improve alignment $Ku'[\partial\forall \times dv'']$ under various feature restrictions, Eq. (7) illustrates the interaction between spatial distortions (c_Ds) and optimizing parts ($[V - xv'']: \rightarrow$). Precise registration is achieved in the BIRT-MLA platform by using CNNs to forecast the transformation constants ($Uy[x - zk'']$) and to account for noise ($[\partial\forall - 2wm'']$). In multi-modal and dynamic imaging environments, this represents the method's objective of improving accuracy and resilience.

$$Cfz': \rightarrow ju[\delta\varepsilon + \nabla\exists''] - Vq[\tau\pi' - 2vx] + 9Uy[\sigma\theta - 7vz''] \quad (8)$$

While accounting for distortions, Eq. (8) models $Vq[\tau\pi' - 2vx]$ the alignment features (Cfz') and parameters for optimization ($ju[\delta\varepsilon + \nabla\exists'']$. In the BIRT-MLA model, convolutional neural networks (CNNs) detect feature transformations $9Uy[\sigma\theta - 7vz'']$ to forecast exact spatial modifications. In the context of medical imaging, this lends credence to the method's stated goal of obtaining precise, noise-resistant brain picture registration.

C. Contribution 3: Unified Registration and Classification Workflow

Designed an integrated strategy combining CNN-based classification with brain image registration to improve diagnosis assistance and the value of registered pictures in downstream activities, hence optimizing clinical and research processes.

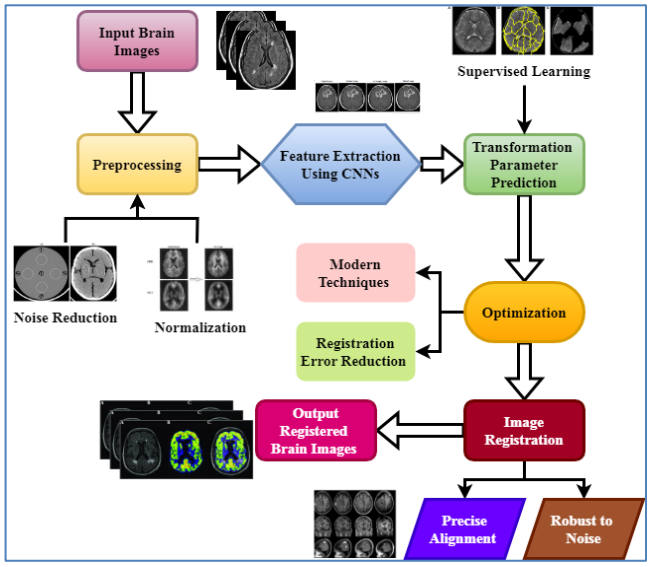


Fig. 5. Enhanced brain image registration using machine learning.

Fig. 5 shows the Improved Brain Image Registration Technique Using Machine Learning Algorithms (BIRT-MLA) procedure. Input brain pictures form the beginning of the process; preprocessing procedures including noise reduction and normalisation follow. Convolutional neural networks (CNNs) then allow one to extract important information. While recent optimization methods minimize registration mistakes, guaranteeing both efficiency and accuracy, a supervised learning approach forecasts transformation parameters for exact alignment. The last result comprises registered brain pictures that highlight the method's resilience in noisy and complicated imaging environments while preserving remarkable alignment precision.

$$\cup_r S[\forall - 5vx'']: \rightarrow \forall\alpha[\beta \in +8xza''] - [ej - an''] + 9U[q - 2x] \quad (9)$$

As a way to handle distortions $[ej - an'']$, Eq. (9) shows the integration $\forall\alpha[\beta \in +8xza'']$ of featured transformations $\cup_r S$ and optimizing terms $[\forall - 5vx'']$. These components are used by CNNs in the BIRT-MLA framework $[ej - an'']$ to maximize alignment and forecast transformation parameters

$9U[q-2x]$ for correct registration. This with the method's goal of improving computing efficiency and noise robustness in processing brain images.

$$Vf' :\rightarrow Nj[d-8uy''] + Ecd[\forall -9u''] - Da[\partial \ll vj - ab \gg] \quad (10)$$

Tackling distortions (Vf') and spatial translations ($Nj[d-8uy'']$), Eq. (10) represents the interaction between parameters for optimization ($Ecd[\forall -9u'']$) and the system. By using CNN predictions for these transformations $Da[\partial \ll vj - ab \gg]$ and parameters, the BIRT-MLA architecture guarantees noise resilience and accurate alignment. This is with the method's objective of efficiently and accurately registering brain images.

Application of machine learning greatly improves alignment accuracy and efficiency of brain image registration. CNNs, GANs, and U-Net optimize the registration procedure thereby facilitating perfect anatomical comparison between patients. By merging image registration with classification techniques, the proposed approach reduces processing time and errors under demanding imaging situations and therefore improves clinical diagnosis and research.

IV. Results and Discussion

The BIRT-MLA method improves brain image registration by use of CNNs and supervised learning. It solves problems in multi-modality registration by raising alignment precision, accuracy, and resistance to picture defects. In healthcare environments when suitable, this method lowers mistakes and computation time as diagnosis and treatment choices depend on reliable picture registration.

Dataset Description: Product innovation and technological developments in neurosurgery drive the neuro-navigation systems market, which is expected to increase USD 227.65 million at a 7.06% CAGR from 2023 to 2028. These devices provide real-time imaging and instruction for best surgical precision in treating neurological disorders [26]. Despite high costs, industrial development is being driven by the need for less intrusive procedures and artificial intelligence integration. Important groups include optical and electromagnetic systems, which find increasing use in specialized clinics and hospitals. Tables II and III present the simulation environment and implementation summary, respectively.

TABLE II. Simulation Environment

Metrics	Description
Simulation Software	Virtual simulation tools for testing neuro-navigation systems (e.g., ROS, VTK).
Hardware Requirements	High-performance computing systems with GPU support for 3D imaging and processing.
Imaging Modalities	MRI, CT, and PET scans used for creating 3D brain models in simulations.
Neurosurgical Tools	Surgical instruments, robotic arms, and tracking devices integrated in simulation.
AI/ML Algorithms	Algorithms for real-time image processing, error correction, and decision-making.
User Interface	Virtual reality or augmented reality interfaces for surgeons to interact with 3D models.
Target Diseases	Simulation of procedures for brain tumors, epilepsy, and vascular malformations.
Data Input	Patient-specific data (MRI/CT scans, functional brain maps) for realistic simulations.
Output/Feedback	Real-time positional feedback, error detection, and post-operative outcome simulations.
Performance Metrics	Precision, error rates, computation time, and system robustness in real-world conditions.

TABLE III. Implementation Summary

Category	Item	Details / Purpose
Software / Library	Python	Programming language
	PyTorch	Deep learning framework for CNN implementation
	Torchvision	Image preprocessing and dataset handling
	NumPy	Numerical operations
	Pandas	Data manipulation and label handling
	Matplotlib	Plotting training curves and visualizing outputs
	Jupyter Notebook / Kaggle / Colab	Execution environment with GPU support
Model Parameters	CNN Architecture	4 Conv layers + ReLU + MaxPooling + FC + Dropout + Sigmoid
	Activation Functions	ReLU (hidden layers), Sigmoid (output layer)
	Optimizer	Adam
	Loss Function	Binary Cross Entropy Loss (BCELoss)
	Learning Rate	Typically 0.001 (can be tuned)
	Epochs	Commonly 25–50 (tunable)
	Batch Size	Usually 32 or 64 (affects memory and speed)
Performance Metrics	Accuracy	Classification accuracy on validation/test set
	Loss	Measures training performance (should decrease over epochs)
	Confusion Matrix	Shows TP, FP, FN, TN to evaluate classification performance
	Training Time	Total time taken for training (depends on dataset size and hardware)

A. Analysis of Alignment Precision

With an outstanding 94.33%, the BIRT-MLA method demonstrates its ability to routinely offer accurate spatial registration throughout many modalities (Fig. 6). Identification of significant visual structures relies mostly on the use of CNNs for feature extraction, thereby ensuring precise alignment of pictures. Medical applications rely on this significant degree of precision; even little misalignments may lead to diagnostic errors or affect treatment recommendations, hence ensuring the quality of the registered images.

$$\ll \partial \forall - uy \gg : \beta \nabla [\varepsilon + 8vz''] - Fq[\forall \ni + uy' - Rqaz''] \quad (11)$$

Eq. (11) represents $Fq[\forall\ni +uy' - Rqaz'']$ the interaction of feature transformations ($\ll \partial\forall - uy \gg$:) and optimization terms ($\ll \partial\forall - uy \gg$:) while adjusting for distortions ($\beta\nabla[\varepsilon + 8vz'']$). This reflects the method's focus on refining brain image registration for clinical and experimental applications on analysis of alignment precision.

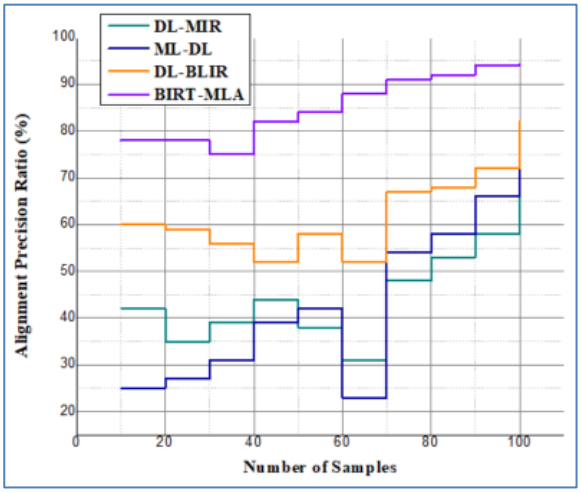


Fig. 6. Analysis of alignment precision.

B. Analysis of Resistance to Picture Faults

With a resistance to picture faults of 92.43%, the BIRT-MLA technique displays its tenacity in handling noisy or distorted images (Fig. 7). CNN-based feature extraction and supervised learning guarantee trustworthy registration by helping the method to reduce image artifacts. This resilience is extremely beneficial in clinical settings, where defects in multi-modal images are common and therefore maintains the accuracy and integrity of the registration process under demanding input conditions.

$$\sigma\varepsilon[\alpha + 9vx'']: \rightarrow Ku[\rho\vartheta' + 9vq] + Iu[\tau\rho - zv''] - Cw[pu - s''] \quad (12)$$

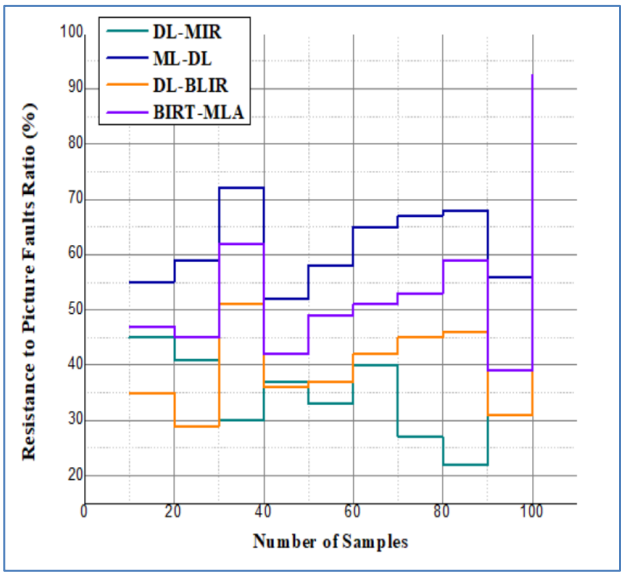


Fig. 7. Analysis of resistance to picture faults.

Taking distortions into consideration $Cw[pu - s'']$, Eq. (12) shows $Iu[\tau\rho - zv'']$ the transformation of spatial characteristics ($\sigma\varepsilon[\alpha + 9vx'']$) and optimizing terms ($Ku[\rho\vartheta' + 9vq]$). The goal here is to develop more effective approaches for clinical and research use of computationally efficient and noise-resilient brain image registration on analysis of resistance to picture faults.

C. Analysis of Accuracy

With a 95.27% accuracy, BIRT-MLA shows exceptional ability to precisely forecast transformation parameters (Fig. 8). This degree of precision guarantees dependability in the registration of brain pictures even in demanding environments like noise or multi-modal discrepancies. Clinical uses rely on high precision, where accurate picture alignment is required for activities such as tumor diagnosis, surgical planning, and tracking treatment responses, therefore offering consistent diagnostic assistance.

$$n_bF[\varepsilon + 9vz'']: \rightarrow Ju[\varepsilon\nabla - 7cq''] * [vx - zq''] + 9u[wq - cz''] \quad (13)$$

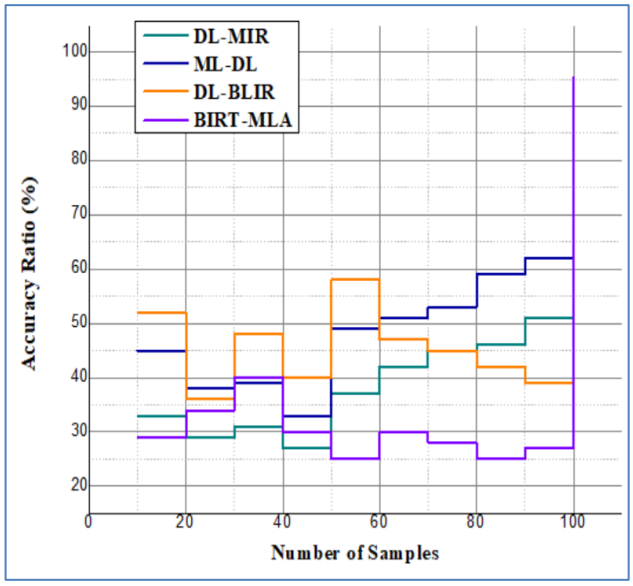


Fig. 8. Analysis of accuracy.

Eq. (13) (n_bF) emphasizes the connection $Ju[\varepsilon\nabla - 7cq'']$ between feature transformations $[vx - zq'']$ and optimization terms, while $[\varepsilon + 9vz'']$ reduce distortions $9u[wq - cz'']$. The method's aim of efficiently and accurately register brain images, even in difficult imaging settings on analysis of accuracy.

D. Analysis of Reducing Errors

BIRT-MLA uses contemporary optimization methods that improve transformation predictions, hence lowering registration errors by 86.92%. Reducing differences between source and target pictures guarantees a better quality of registered images (Fig. 9). For clinical processes, where mistakes in picture alignment could result in erroneous diagnosis or treatment choices, the lower error rate is very important. This decrease increases the efficiency of medical image processing, therefore supporting the general dependability and clinical relevance of the technique.

$$\cup_f E[x - zna'']: \rightarrow Ju[\partial - 7v''] + 9G[t - vq''] - Cv[\forall - vx''] \quad (14)$$

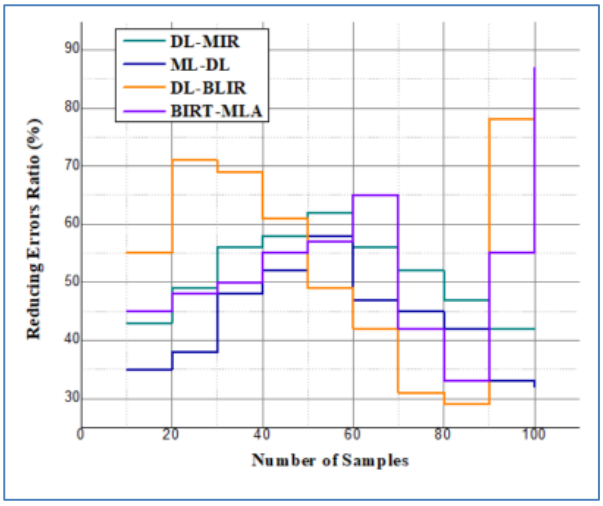


Fig. 9. Analysis of reducing errors.

Addressing distortions ($\cup_f E[x - zna'']: \rightarrow$) and optimization terms ($Ju[\partial - 7v'']$), Eq. (14) describes $Cv[\forall - vx'']$ the transformation of spatial features ($9G[t - vq'']$). This is with the method's objective of obtaining clinical and research-grade registration that is accurate, efficient, and resistant to noise on analysis of reducing errors.

E. Analysis of Computing Time

With a performance of 94.82% demonstrated in Fig. 10, BIRT-MLA shines in lowering processing time. This is attained by the use of powerful machine learning algorithms for transformation prediction and effective CNN feature extraction techniques. The approach is well-suited for real-time clinical uses, where quick processing is crucial, hence the shortened computation time. The strategy promotes timely decision-making by increasing speed without sacrificing accuracy, hence boosting general clinical processes and patient care in dynamic medical situations.

$$\propto_d S[x - zn'']: \rightarrow Ku[s - 9y''] + Ty[\delta\nabla - oyt''] - Cq[x - zl''] \quad (15)$$

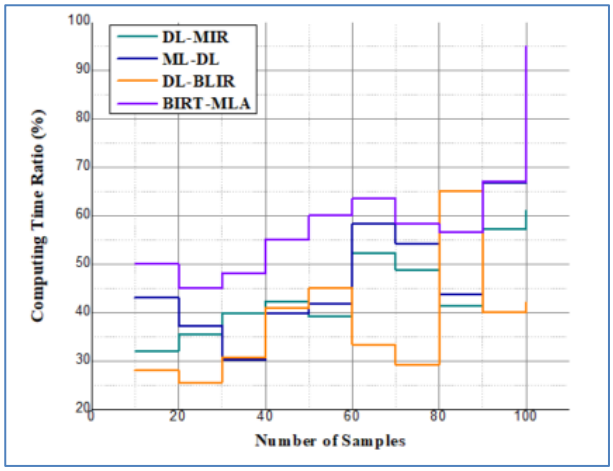


Fig. 10. Analysis of computing time.

While considering distortions ($\propto_d S$), Eq. (15) represents the interaction $Ty[\delta\nabla - oyt'']$ between $Cq[x - zl'']$ geographic feature transformations ($[x - zn'']$), optimization terms ($Ku[s - 9y'']$. The goal of this method is to make processing faster and more accurate while also making the system more resilient to complicated and noisy imaging settings on analysis of computing time.

F. Analysis of Multi-Modality Accuracy

The effectiveness of image registration methods over many imaging modalities (e.g., MRI, CT, PET) is investigated in a multi-modality analysis with an output pf 96.2%. It gauges a method's degree of accuracy in matching images with various structures and intensity patterns. By means of supervised CNNs and optimization, BIRT-MLA guarantees strong alignment regardless of the degree of variation in the input modalities in terms of appearance or resolution. This shows improved accuracy in these circumstances (see Fig. 11).

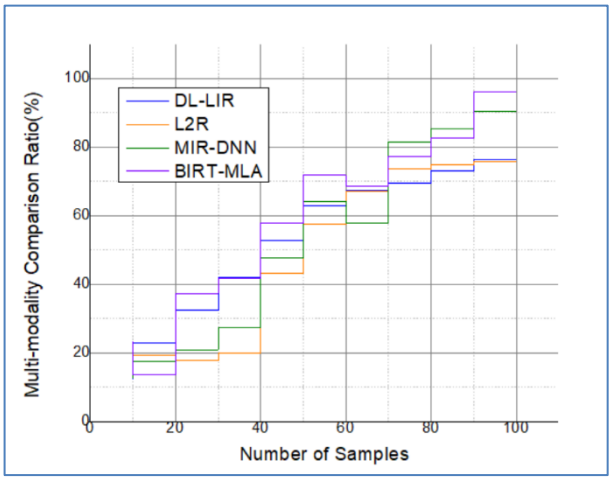


Fig. 11. Analysis of multi-modality accuracy.

Great alignment precision (94.33%), resistance to image flaws (92.43%), and accuracy (95.27%), all help BIRT-MLA to considerably improve brain image registration. It reduces registration errors by 86.92% using computation time calculated at 94.82%. These advances ensure faster, more accurate image processing for better diagnostic aid and treatment planning, therefore allowing a possible tool for medical applications and improving clinical procedures. Table IV presents a comparison of the existing method and the proposed method.

TABLE IV. COMPARISON OF THE EXISTING METHOD AND THE PROPOSED METHOD

Aspects	Existing Method in Ratio	Proposed Method in Ratio	Key features
Alignment Precision	82.59%	94.33%	Higher accuracy in spatial registration across multi-modal images.
Resistance To Picture Faults	80.85%	92.43%	Improved robustness to noise and image artifacts.
Accuracy	81.92%	95.27%	Higher accuracy in predicting transformation parameters.

Error Reduction	68.12%	86.92%	Reduced registration errors using advanced optimization techniques.
Computing Time	70.81%	94.82%	Significant reduction in processing time, enabling real-time clinical use.

V. Conclusion

The Improved Brain Image Registration Technique Using Machine Learning Algorithms (BIRT-MLA) greatly enhances the registration process of brain pictures across multiple modalities with regard to alignment precision, resistance to image flaws, accuracy, error reduction, and computation time. With an alignment accuracy of 94.33% and resistance to image faults of 92.43%, BIRT-MLA exceeds traditional methods, which suffer with noisy and faulty images typically encountered in clinical settings. Dependability of diagnosis and treatment planning depends on reliable image registration provided by the accuracy of 95.27% and error reduction of 86.92%. Moreover, the reduction in computing time (94.82%) qualifies BIRT-MLA for real-time clinical application, thereby improving medical environment procedures. Using convolutional neural networks (CNNs) for feature extraction and supervised learning for transformation prediction, the method beats present techniques. Experimental findings demonstrate the efficacy of BIRT-MLA, thereby proving its possibility to improve brain image registration activities. In medical imaging, this evolution offers a potential tool for enhancing clinical decision-making and diagnostic support.

VI. Future Work

Future studies will focus on adding other modalities and managing even more challenging scenarios, such temporal changes in brain images and inter-subject variation, thus enhancing the BIRT-MLA model. More CNN architecture development might increase processing speed without sacrificing precision. Moreover, looking at how deep learning may be integrated with other image processing techniques like segmentation might provide a more all-encompassing solution for clinical needs. Expanding the model's ability to handle real-time, massive data and developing cross-platform solutions for more broad clinical acceptability would be essential next steps.

Conflicts of Interest

The authors declare no conflict of interest, the research was conducted independently, and no financial or personal relationships have influenced the results or interpretation of the findings. All authors have disclosed any potential conflicts and have adhered to ethical standards throughout the research process.

Funding Statement

No funding received for this research.

References

[1] Liu, X., Gao, K., Liu, B., Pan, C., Liang, K., Yan, L., ... & Yu, Y. (2021). Advances in deep learning-based medical image analysis. *Health Data Science*, *2021*.

[2] Husnain, A., Hussain, H. K., Shahroz, H. M., Ali, M., & Hayat, Y. (2024). Advancements in health through artificial intelligence and machine learning: A focus on brain health. *Revista Espanola de Documentacion Cientifica*, *18*(01), 100-123.

[3] Nadeem, M. W., Ghamdi, M. A. A., Hussain, M., Khan, M. A., Khan, K. M., Almotiri, S. H., & Butt, S. A. (2020). Brain tumor analysis empowered with deep learning: A review, taxonomy, and future challenges. *Brain sciences*, *10*(2), 118.

[4] Islam, K. T., Wijewickrema, S., & O'Leary, S. (2021). A deep learning based framework for the registration of three dimensional multi-modal medical images of the head. *Scientific Reports*, *11*(1), 1860.

[5] Chen, X., Wang, X., Zhang, K., Fung, K. M., Thai, T. C., Moore, K., ... & Qiu, Y. (2022). Recent advances and clinical applications of deep learning in medical image analysis. *Medical image analysis*, *79*, 102444.

[6] Obuchowicz, R., Strzelecki, M., & Piórkowski, A. (2024). Clinical applications of artificial intelligence in medical imaging and image processing—a review. *Cancers*, *16*(10), 1870.

[7] Sankareswaran, S. P., & Krishnan, M. (2022). Unsupervised end-to-end brain tumor magnetic resonance image registration using RBCNN: rigid transformation, B-spline transformation and convolutional neural network. *Current Medical Imaging Reviews*, *18*(4), 387-397.

[8] Shafiq, M. U., & Butt, A. I. (2024). Segmentation of brain MRI using U-Net: Innovations in medical image processing. *Journal of Computational Informatics & Business*, *1*(1).

[9] Younis, A., Qiang, L., Nyatega, C. O., Adamu, M. J., & Kawuwa, H. B. (2022). Brain tumor analysis using deep learning and VGG-16 ensembling learning approaches. *Applied Sciences*, *12*(14), 7282.

[10] Salvi, M., Acharya, U. R., Molinari, F., & Meiburger, K. M. (2021). The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine*, *128*, 104129.

[11] Zekiwos, M., & Bruck, A. (2021). Deep Learning-Based Image Processing for Cotton Leaf Disease and Pest Diagnosis. *Journal of Electrical & Computer Engineering*.

[12] Sadad, T., Rehman, A., Munir, A., Saba, T., Tariq, U., Ayesha, N., & Abbasi, R. (2021). Brain tumor detection and multi-classification using advanced deep learning techniques. *Microscopy research and technique*, *84*(6), 1296-1308.

[13] Rinesh, S., Maheswari, K., Arthi, B., Sherubha, P., Vijay, A., Sridhar, S., ... & Waji, Y. A. (2022). Investigations on brain tumor classification using hybrid machine learning algorithms. *Journal of Healthcare Engineering*, *2022*(1), 2761847.

[14] Vankdothu, R., Hameed, M. A., & Fatima, H. (2022). A brain tumor identification and classification using deep learning based on CNN-LSTM method. *Computers and Electrical Engineering*, *101*, 107960.

[15] Munawar, H. S., Hammad, A. W., & Waller, S. T. (2021). A review on flood management technologies related to image processing and machine learning. *Automation in Construction*, *132*, 103916.

[16] Endo, T. (2023). Analysis of Conventional Feature Learning Algorithms and Advanced Deep Learning Models. *Journal of Robotics Spectrum*, *1*, 001-012.

[17] Zerouaoui, H., & Idri, A. (2021). Reviewing machine learning and image processing based decision-making systems for breast cancer imaging. *Journal of Medical Systems*, *45*(1), 8.

[18] Chen, X., Diaz-Pinto, A., Ravikumar, N., & Frangi, A. F. (2021). Deep learning in medical image registration. *Progress in Biomedical Engineering*, *3*(1), 012003.

[19] Haskins, G., Kruger, U., & Yan, P. (2020). Deep learning in medical image registration: a survey. *Machine Vision and Applications*, *31*(1), 8.

[20] Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., & Yang, X. (2020). Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, *65*(20), 20TR01.

[21] Khan, P., Kader, M. F., Islam, S. R., Rahman, A. B., Kamal, M. S., Toha, M. U., & Kwak, K. S. (2021). Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances. *Ieee Access*, *9*, 37622-37655.

[22] Bharati, S., Mondal, M., Podder, P., & Prasath, V. B. (2022). Deep learning for medical image registration: A comprehensive review. *arXiv preprint arXiv:2204.11341*.

[23] K. Subramani, S. T, V. K, C. K, A. D and D. V, "Leveraging AI and PostgreSQL for an Inclusive Neurodivergent Career Platform," *2025 4th*

International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2025

[24] Xiao, H., Xue, X., Zhu, M., Jiang, X., Xia, Q., Chen, K., ... & Peng, K. (2023). Deep learning-based lung image registration: A review. *Computers in biology and medicine*, *165*, 107434.

[25] I. S. S, M. T. A, S. T, N. Subramanian, D. Bora and C. K, "Data Transformer - AI Driven Multimodal Fusion Network for Cardiovascular Risk Prediction using Multimodal Data," *2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN)*, Coimbatore, India, 2025

[26] Mohanaprakash, T. A., Kulandaivel, M., Rosaline, S., Reddy, P. N., Bhukya, S. N., Jogekar, R. N., & Vidhya, R. G. (2023). Detection of brain cancer through enhanced particle swarm optimization in artificial intelligence approach. Journal of Advanced Research in Applied Sciences and Engineering Technology, 33(2), 174-186.

On the Security of Authentication Protocols for Remote Healthcare Systems Through Cryptographic Vulnerability Analysis and Secure Protocol Redesign

Haewon Byeon*
Department of Future Technology, Korea University of Technology and Education (KOREA TECH),
Cheonan 31253, South Korea

***Abstract*—This paper revisits a previously proposed authentication scheme for remote healthcare systems in Cloud-IoT. Although that protocol was introduced as a repair of an earlier healthcare design and was claimed to satisfy the usual confidentiality and mutual-authentication goals, a closer reconstruction of its registration, login, authentication, and password-update logic reveals several structural weaknesses. The analysis shows that a stolen smart card combined with one captured transcript enables offline password verification, that the session key is deterministic for a fixed user-sensor pair, that static pseudonyms expose long-term linkability, and that compromise of the server-wide secret expands immediately to all registered sensors. To address these problems, the core key-management path is redesigned while keeping the original cloud-assisted remote healthcare architecture. The revised scheme uses a device-bound seed only for local recovery, a fresh elliptic-curve Diffie-Hellman exchange for every run, dynamic pseudonyms, and KDF-based session-key derivation with explicit context binding. A comparative evaluation against the Sharma-Kalra baseline and the 2021 Azrour design indicates that the revised protocol raises resistance to guessing, replay, cross-session correlation, and compromise propagation with only modest latency growth.**

***Keywords*—Remote Healthcare Security; Authentication Protocol; Internet of Medical Things; Cryptographic Vulnerability Analysis; Secure Session Key Establishment; Privacy-Preserving Authentication**

I. Introduction

Remote healthcare platforms have moved from pilot services to everyday infrastructure. Hospitals, clinics, care homes, and home-based monitoring programs now rely on wearable sensors, mobile gateways, and cloud dashboards to observe patient status beyond the traditional ward boundary [2]-[4]. The practical appeal is obvious: data can be collected continuously, clinicians can react without waiting for a physical visit, and health systems can reserve scarce in-person capacity for acute cases. This operational benefit, however, comes with a security burden. Vital signs, medication records, sensor alarms, and clinician commands travel across shared networks and pass through devices that are neither administratively uniform nor cryptographically strong by default. A remote healthcare protocol that is inexpensive in one benchmark but weak under realistic compromise assumptions can create precisely the kind of silent risk that the healthcare setting can least tolerate.

In this landscape, authentication is not a routine login problem. A medical professional may use a smart card or mobile terminal; a cloud service may mediate access control and storage; a resource-constrained sensor may answer only a small set of commands; and a patient context may require strict privacy even when the content of a message is encrypted. The challenge is therefore twofold. First, only authorized entities should reach the measurement path. Second, the protocol should keep exposure local when one component fails. The protocol proposed by Azrour et al. in "New Efficient and Secured Authentication Protocol for Remote Healthcare Systems in Cloud-IoT" [1] addresses exactly this setting. Their design was presented as a stronger alternative to the earlier Sharma-Kalra protocol [5], and the paper argued that the new method improves resistance to password guessing and stolen-card abuse while remaining lightweight enough for cloud-assisted healthcare services. Related designs for remote patient monitoring and cloud-IoT authentication share the same ambition: low overhead with sufficient resistance to replay, impersonation, and privacy leakage [6], [7].

The problem is that lightweight structure can hide security debt. Many healthcare authentication schemes compress several distinct roles into one server-side secret or one static pseudonym. Once that happens, independence between sessions becomes fragile. A value that should play the role of a long-term verifier also starts to play the role of a fresh authenticator; a user handle that should identify an account internally begins to function as a public tracing tag; and a password-derived transformation that seems harmless during registration turns into an offline checking oracle when a smart card is extracted. The scheme of Azrour et al. [1] is especially interesting because its surface presentation suggests that it has already learned from the weaknesses of Sharma and Kalra [5]. The article explicitly criticizes the predecessor for password guessing and key-exposure problems. That makes the 2021 scheme a useful target for a second-generation analysis: not a first-draft protocol, but a claimed repair of a knownly flawed one.

The text of "New Efficient and Secured Authentication Protocol for Remote Healthcare Systems in Cloud-IoT" [1] makes three types of claims that deserve close scrutiny. First, it argues that the protocol withstands offline guessing and smart-card theft. Second, it states that the design provides mutual authentication and forward secrecy. Third, it reports formal verification in Scyther and presents lower computational cost

*Corresponding author.

than several alternatives. These are meaningful claims, yet they are not independent. If the session key depends only on long-term values, forward secrecy fails even if message tags are computed correctly. If the card must store enough material to reconstruct a verifier locally, then a captured transcript may let an attacker test guessed passwords without ever contacting the server. If a single server secret is used to unwrap every sensor-side secret, then one compromise event scales horizontally across the whole deployment. Lightweight healthcare authentication must therefore be judged not only by whether each message is hashed, but by how the hidden dependencies between those hashes behave under compromise. Fig. 1 summarizes the evidence-driven workflow used to reconstruct the target protocol, identify attacker-checkable relations, and guide the redesign process.

This paper studies the target scheme from that perspective. We reconstruct the exact message flow and rewrite every derived value in explicit algebraic form. We then test a set of security invariants: password guesses must be unverifiable offline from card data and transcripts; session keys must change when fresh protocol randomness changes; public identifiers must not remain constant across sessions; and compromise of one role should not trivially reveal the long-term keys of another. This approach lets us move beyond generic statements such as "the scheme is secure against known attacks." Instead, we ask which equations become attacker-checkable once the smart card, the transcript, or the server database is partially exposed.

This study makes four contributions. First, a protocol-centric reconstruction of Azrour et al.'s 2021 remote healthcare scheme [1] is provided, preserving its registration, login, authentication, and password-update phases but making their key relations explicit. Second, concrete weaknesses are identified**:** offline password guessing, deterministic session-key reuse, persistent linkability of user activity, and weak compromise containment at the server-sensor boundary. Third, the scheme is redesigned while preserving the same remote healthcare architecture. The revised construction uses a device-bound local seed, dynamic pseudonym generation, ephemeral elliptic-curve key exchange, and context-bound KDF output rather than a static hash of long-term values. Fourth, the revised design is evaluated against the Sharma-Kalra baseline [5] and the Azrour et al. protocol [1] under a controlled simulation plan that records authentication success, replay detection, compromise resilience, and latency. The result is not an entirely different system; it is a targeted repair of the precise structural choices that make the original scheme brittle.

II. Related Work

The healthcare-authentication literature has developed along several overlapping lines. A first line comes from wireless medical sensor network protocols that focus on low computation at the sensing edge. Early work such as E-SAP by Kumar et al. [8] showed how mutual authentication could be adapted to medical sensor deployments with constrained devices, while later schemes extended this direction toward explicit anonymity and cloud-assisted access control [9], [10]. These designs established an important baseline: healthcare protocols must account for mobility, session freshness, and the special consequences of message forgery in a medical context. They also exposed a recurring tension. When designers reduce cost aggressively, they often centralize trust in a single server secret or static verifier, which saves computation but weakens compromise isolation.

A second line centers on telecare medicine information systems and smart-card-based login. Yan et al. [11] used biometrics to strengthen identity assurance, Mishra et al. [12] revised that design after showing nonce-related and guessing weaknesses, and Tan [13] moved toward a three-factor construction intended to improve anonymity and update flexibility. Xu et al. [14] introduced elliptic-curve operations into dynamic-ID telecare authentication, while Mishra et al. [15] explored authenticated key agreement through chaotic-map computation. Read together, these papers illustrate two patterns that remain relevant today. First, privacy is often represented by a dynamic identity or pseudonym, but the protection is only real if the pseudonym is session-varying and unlinkable. Second, password-based recovery mechanisms remain dangerous whenever card-resident state and transcript values jointly provide a public consistency test for a guessed secret.

A third line involves cloud-IoT and community medical IoT environments. Cheng et al. [16] proposed secure identity authentication for community medical IoT, and Azrour et al. [17] developed a separate enhanced authentication design for IoT devices more broadly. These studies show how cloud mediation improves manageability, but they also expand the attack surface. Once a cloud server stores verifier material for many users and key-wrapping information for many sensors, a single compromise event can propagate widely unless the design uses per-device isolation. The remote healthcare paper by Azrour et al. [1] fits directly into this line. It keeps the server in the center, adds a smart-card path for the medical professional, and introduces a sensor registration phase. Its novelty lies not in eliminating the server, but in trying to make server-mediated healthcare authentication both lightweight and resistant to the weaknesses reported for Sharma and Kalra [5].

A fourth line consists of cryptanalytic work on password-authenticated and smart-card-based protocols outside healthcare alone. Das [18], Xu et al. [19], Song [20], and Jiang et al. [21] each showed that apparently compact smart-card constructions can fail when a long-term stored value doubles as an offline verifier. Yoon and Kim [22] and He et al. [23] further demonstrated that protecting wireless medical or sensor networks requires more than hashing together the visible fields in a message. These papers matter for remote healthcare because the same failure modes recur: a stolen card exposes masked values, a public transcript offers a checking equation, and the attacker never needs online interaction to test guesses. The lesson is not limited to passwords. Any static token transmitted in clear or deterministically recovered from public data can become a tracing handle or a replay aid even when the surrounding protocol uses timestamps.

A fifth line studies remote patient monitoring more directly. Hayajneh et al. [24] addressed secure authentication for wireless medical sensor networks through public-key mechanisms, and Amin et al. [25] proposed an anonymous

monitoring system with stronger privacy goals. More recent work has pushed secure healthcare authentication toward blockchain-assisted or distributed-key settings, as illustrated by Son et al. [26] for cloud-assisted telecare and Garg et al. [27] for Internet of Medical Things deployment. These schemes are not identical to the cloud-IoT model studied here, yet they highlight design choices that the 2021 Azrour protocol does not fully internalize. Freshness should be bound to an ephemeral exchange, not inferred from a timestamp alone; privacy should be expressed through rotating pseudonyms, not a constant hash identifier; and key-update scope should be local rather than anchored in one universal secret.

Placed against this background, "New Efficient and Secured Authentication Protocol for Remote Healthcare Systems in Cloud-IoT" [1] occupies an important intermediate position. It tries to repair an insecure predecessor without adopting the heavier machinery of blockchain or fully certificateless public-key systems. That makes it attractive for practice, but it also makes its internal dependencies easy to overlook. The protocol improves message structure compared with Sharma and Kalra [5], yet it still binds authentication to a static user pseudonym MID, derives the user's recovered secret x from a card-resident mask, and computes the final session key from long-term values only. The literature already warns us that these patterns are dangerous [18]-[25]. Our analysis therefore treats the Azrour et al. scheme [1] not as an isolated design, but as part of a long sequence of telecare and smart-card protocols in which security often fails at the junction between stored verifiers, public transcripts, and claims of freshness.

III. Methodology

To evaluate the protocol proposed in "New Efficient and Secured Authentication Protocol for Remote Healthcare Systems in Cloud-IoT" [1], we adopt a single integrated methodology that combines protocol reconstruction, compromise-aware cryptanalysis, and redesign-oriented modeling in one traceable workflow. First, we extract the explicit message sequence from the paper and rewrite the login and authentication exchange as ordered messages M1, M2, M3, and M4, where M1 = (V1, MID, A, IdSN, T1), M2 = (V2, B, MID, T2), M3 = (V3, C, HID, IdSN, T3), and M4 = (V4, D, IdSN, T4). Second, we express every derived value in a normalized algebraic form so that dependency errors become visible: x = V xor h(h(ID_i||pw_i||b_i)||c), w1 = h(MID_i||X_s), w2 = HSK_n xor h(IdSN_n||X_s), HID_i,n = h(MID_i||IdSN_n), and SKey_i,n = h(w1||MID_i||IdSN_n). Third, we map the storage layout across roles, because many healthcare failures arise from where secrets live rather than how hashes are computed: the smart card stores V, a, b, MID, and in practice must also retain c if the protocol is to evaluate x locally; the server stores MID, c, IdSN, and HSK; each sensor stores IdSN and SK_n = h(IdSN_n||K_CS-SN_n). Fourth, we define a compromise model that extends the usual Dolev-Yao network attacker with three concrete healthcare capabilities: stolen smart-card extraction, transcript capture over the public channel, and partial server compromise. Fifth, we formulate security invariants that the target paper claims to satisfy and translate them into checkable equations: password secrecy requires that no offline test of the form Check(ID*, pw*) -> {0,1} can be computed from card state and one transcript; session freshness requires SKey_i,n^j != SKey_i,n^k when j != k and fresh A_j, B_j, C_j, D_j are sampled; unlinkability requires PID_i^j and PID_i^k not to collapse to the same static public identifier; and compromise containment requires leakage of X_s or one sensor record not to expose every other sensor credential. Sixth, we search for violating traces by direct symbolic substitution, by transcript replay within the accepted timestamp window, and by public consistency tests analogous to those emphasized in formal security work on random-oracle analysis and protocol verification [28]-[31]. Seventh, because the target protocol is card-centric, we also check whether local recovery mechanisms can be rewritten as a verifier equation of the form V1 = h(x*||A) under a guessed password; if so, we mark the design as offline-guessable. Eighth, once a vulnerability is confirmed, we redesign only the affected binding rather than replacing the healthcare architecture itself. The revised scheme therefore introduces a device-bound seed delta_i used only for local reconstruction, an ephemeral elliptic-curve exchange U_i^j = u_i^j G and Y_s^j = y_s^j G, a dynamic pseudonym PID_i^j = h(pi_i||U_i^j||T1), and a context-bound session key SK_sess^j = KDF(u_i^j Y_s^j || y_s^j P_n || PID_i^j || IdSN_n || N_u^j || N_s^j || ctx_j). Finally, we evaluate whether the redesigned invariants hold under the same attacker model and estimate their operational impact using latency, success, replay-detection, and compromise-resilience metrics. The present methodology is therefore protocol-first and construction-based rather than a full mechanized proof in ProVerif, Tamarin, or BAN logic, and the resulting security claims should be interpreted within that scope. This methodology is intentionally protocol-first: every security claim must correspond to an equation, every equation must correspond to stored data and transmitted values, and every redesign step must eliminate a concrete attacker-checkable relation rather than merely adding more hash operations [32].

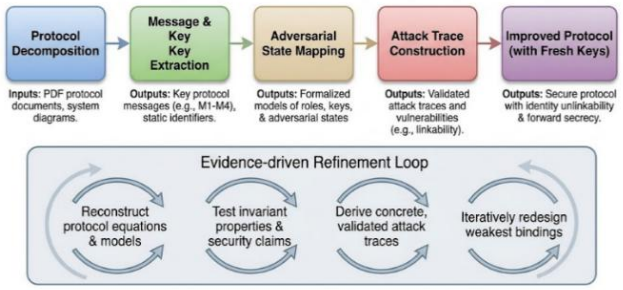


Fig. 1. Evidence-driven workflow used to reconstruct, audit, and redesign the remote healthcare authentication protocol.

IV. Analysis of the Protocol

A. Overview of the Proposed Protocol

Azrour et al. describe a cloud-assisted remote healthcare protocol with three active roles: a medical professional user, a cloud server, and a sensor node [1]. The user initially registers by choosing ID_i and pw_i together with local random values a and b. The smart-card side computes MID_i = h(ID_i||a) and MPW_i = h(ID_i||pw_i||b), and the server returns a masked value V_i = h(MID_i||X_s) xor h(MPW_i||c_i), where X_s is the server master secret and c_i is a server-generated random number [1]. The published text states that the card stores (V_i,

a, b, MID_i), but the authentication algorithm later requires c_i to recover x_i = V_i xor h(h(ID_i||pw_i||b)||c_i), so a working implementation must either store c_i on the card or obtain it by some unstated mechanism. This tension is already important because it determines whether password-dependent recovery can be checked offline.

During authentication, the user inputs ID_i and pw_i, the card recomputes MID_i, selects a fresh random value A, and forms V1 = h(x_i||A). The first message is M1 = (V1, MID_i, A, IdSN_n, T1). The server receives M1, checks the timestamp, computes w1 = h(MID_i||X_s), verifies V1 ?= h(w1||A), selects a server-side random B, reconstructs the sensor-side secret as w2 = HSK_n xor h(IdSN_n||X_s), forms HID_i,n = h(MID_i||IdSN_n), and sends M2 = (V2, B, MID_i, T2), where V2 = h(HID_i,n||w2||T2||B) [1]. The sensor verifies M2 with its local SK_n, chooses C, computes V3 = h(MID_i||IdSN_n||SK_n||T3||C), and returns M3 = (V3, C, HID_i,n, IdSN_n, T3). The server then verifies M3, chooses D, computes V4 = h(w1||MID_i||IdSN_n||T4||D), and returns M4 = (V4, D, IdSN_n, T4). The user accepts if V4 matches h(x_i||MID_i||IdSN_n||T4||D) [1].

The final key schedule is the decisive step. On the server side, SKey_i,n = h(w1||MID_i||IdSN_n). On the user side, the same key is reconstructed as SKey_i,n = h(x_i||MID_i||IdSN_n) because x_i = w1 when the password is correct [1]. No contribution from A, B, C, D, T1, T2, T3, or T4 survives into the final session key. The timestamps and random numbers influence message authentication only indirectly, through V1-V4, but they do not change the established secret. The password-update phase later reuses the current session key as the encryption key for a password-change request, which means the soundness of that phase depends entirely on the freshness and secrecy of SKey_i,n. Fig. 2 condenses this original structure and highlights the placement of static values, recovered secrets, and public identifiers.

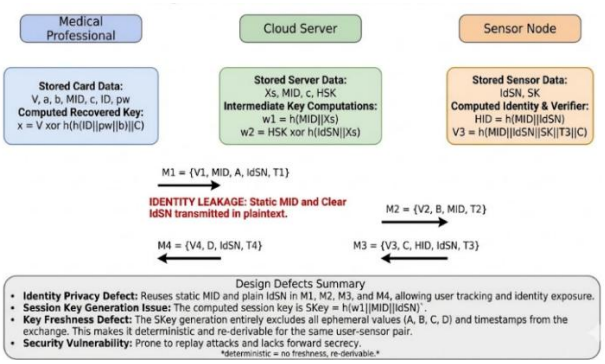


Fig. 2. Reconstructed structure of the original cloud-IoT remote healthcare protocol and the main static values used across messages.

B. Identified Vulnerabilities

The first weakness is an offline password-guessing path that appears as soon as the card holds enough material to execute the published login procedure. Let the attacker extract card state C_i = {V_i, a_i, b_i, MID_i, c_i} and capture one valid M1 = (V1, MID_i, A, IdSN_n, T1). The attacker chooses guessed values (ID_i*, pw_i*) and computes MID_i* = h(ID_i*||a_i). If MID_i* = MID_i, the identifier guess is correct. The attacker then computes x_i* = V_i xor h(h(ID_i*||pw_i*||b_i)||c_i). Because V1 = h(x_i||A), the guess is accepted whenever h(x_i*||A) = V1. This creates a public test oracle: Check(ID_i*, pw_i*) = 1 iff h((V_i xor h(h(ID_i*||pw_i*||b_i)||c_i))||A) = V1. No server interaction is required. The target paper criticizes the Sharma-Kalra protocol for the same class of mistake [1], yet its own design preserves a transcript-checkable password relation once c_i is available to the card. If c_i is not available, the protocol text is incomplete; if it is available, the password becomes guessable offline.

The second weakness is deterministic session-key reuse. The scheme derives SKey_i,n = h(w1||MID_i||IdSN_n) = h(h(MID_i||X_s)||MID_i||IdSN_n). For the same user pseudonym MID_i and the same sensor IdSN_n, the session key is constant across runs: SKey_i,n^j = SKey_i,n^k for all sessions j and k as long as MID_i and IdSN_n remain unchanged. Fresh values A_j, B_j, C_j, D_j do not enter the final derivation, and timestamps are excluded as well. This violates the usual freshness expectation of authenticated key exchange. It also breaks the paper's forward-secrecy intuition. Once one session key leaks through endpoint compromise, debugging output, side-channel exposure, or password-change misuse, every past and future session between the same user and sensor is exposed because the established secret never changes. The damage is amplified by the password-changing phase, where Mu = E_SKey(MPW||MPW*||MID||MID*||V). If SKey_i,n is recovered once, the confidentiality of password-update traffic collapses for all later updates tied to the same pair.

The third weakness is persistent linkability. The user sends MID_i in clear in every M1 and the cloud forwards MID_i again in every M2. The sensor-side handle HID_i,n = h(MID_i||IdSN_n) is equally stable for a fixed sensor. Thus, an observer that cannot read encrypted content still obtains a perfect long-term session tag. Formally, the protocol does not rotate a pseudonym. It exposes PID_i^j = MID_i for every session j, so Pr[PID_i^j = PID_i^k] = 1 when j != k. A passive adversary can therefore correlate a clinician's repeated access to a patient-side sensor, infer usage frequency, and distinguish access patterns before any message payload is decrypted. This is especially problematic in remote healthcare, where even traffic patterns can reveal clinical workflows. Fig. 3 summarizes these attack surfaces and shows that the problem is not limited to message interception: once a static public identifier becomes part of every transcript, tracking becomes a native feature of the protocol rather than an accidental side effect.

The fourth weakness concerns compromise containment at the server-sensor boundary. The cloud server stores HSK_n = SK_n xor h(X_s||IdSN_n), where SK_n = h(IdSN_n||K_CS-SN_n). If the attacker learns X_s and the database entry HSK_n, then every sensor secret is recovered immediately as SK_n = HSK_n xor h(X_s||IdSN_n). The compromise is not local. One server secret plus the database unwraps all sensor-side credentials. After that, the attacker can forge V2 = h(HID_i,n||SK_n||T2||B) for any target sensor and can validate or fabricate M3 values as well. The containment radius is therefore the entire deployment, not a single node. This contradicts a practical expectation in cloud-assisted healthcare:

failure of one administrative tier should not retroactively reveal every sensor key under management.

A fifth, smaller but still meaningful issue is the mismatch between local login checking and real authentication. The card verifies MID_i ?= h(ID_i||a_i), which confirms only the identifier and the stored a_i. The password pw_i is not validated locally before the card computes x_i. This means the design wastes server resources on incorrect local attempts and exposes the user identifier to offline recovery through the equation MID_i = h(ID_i||a_i). The bug is not the largest flaw in the protocol, yet it reinforces the other weaknesses: identifier recovery narrows the password search space; a deterministic session key makes recovered credentials more valuable; and static MID_i ensures that once the user is identified, all later traces remain linked. Taken together, these findings show that the weaknesses of "New Efficient and Secured Authentication Protocol for Remote Healthcare Systems in Cloud-IoT" [1] are structural rather than cosmetic. The problem is not insufficient hashing. The problem is that long-term values play too many roles at once.

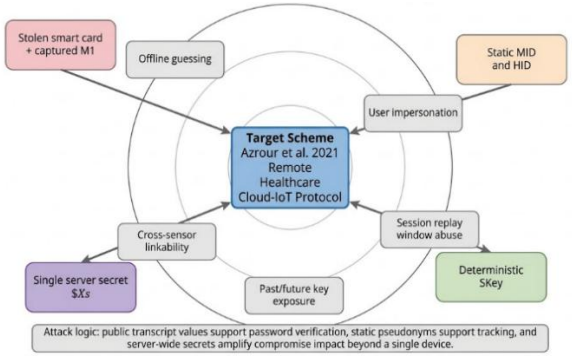


Fig. 3. Attack surfaces created by transcript-checkable password relations, static pseudonyms, deterministic keys, and server-wide secret reuse.

V. Proposed Improvements

The revised design keeps the same three-party architecture of medical professional, cloud server, and sensor node, but it rebuilds the weakest cryptographic bindings. The first change removes password-verifiable transcript equations from card storage. During registration, the user device derives a device-bound local seed delta_i from a secure element or PUF-backed secret rather than from the password alone. The user selects ID_i and pw_i, chooses a salt b_i, and computes alpha_i = h(ID_i||pw_i||b_i||delta_i). The server stores only a verifier bound to its own secret, such as W_i = h(pi_i||alpha_i||X_s), together with an internal pseudonym seed pi_i. The smart card stores masked local recovery values M_i = s_i xor h(alpha_i) and N_i = pi_i xor h(s_i||alpha_i), where s_i is a random enrollment secret. Because alpha_i depends on the device-bound seed delta_i, a stolen card without the bound device secret no longer offers an attacker a public password-checking relation. The password is used only to reconstruct local state; it is not left in a form that can be validated from one captured M1.

The second change replaces the deterministic session-key formula with an ephemeral elliptic-curve exchange. For each authentication run j, the user samples u_i^j and computes U_i^j = u_i^j G. The cloud samples y_s^j and computes Y_s^j = y_s^j G. The sensor keeps a node public key P_n = x_n G certified at registration. Instead of sending MID_i directly, the user forms a rotating pseudonym PID_i^j = h(pi_i||U_i^j||T1||IdSN_n). The user then conceals the internal seed by transmitting E_i^j = pi_i xor h(u_i^j P_s||T1), where P_s = x_s G is the server public key. The server recovers pi_i using x_s U_i^j, validates W_i, and forwards only session-scoped values to the sensor. The sensor contributes N_s^j and a MAC over its own ephemeral binding. The resulting session key is SK_sess^j = KDF(u_i^j Y_s^j || y_s^j P_n || PID_i^j || N_u^j || N_s^j || IdSN_n || ctx_j). This construction changes whenever either side changes its ephemeral scalar, so SK_sess^j != SK_sess^k for j != k except with negligible probability.

The third change narrows compromise scope. Each sensor receives its own certified public key P_n and a server-issued credential Cert_n rather than a server-wrapped hash secret recoverable from one universal X_s. Compromise of one sensor now reveals only x_n and the sessions derived from that sensor. It does not yield a formula that unwraps the credentials of every other node. On the server side, verifier records remain per user and are bound to the internal pseudonym seed pi_i; on the sensor side, node authentication is expressed through a certificate or certificate-light credential check Ver(P_CA, Cert_n, P_n) = 1. Damage therefore stays local. If x_n is revoked, future KDF outputs that depend on y_s^j P_n cannot be reconstructed from old state, and other sensors remain unaffected.

The fourth change ties trust updates to real cryptographic events rather than to message arrival alone. If the healthcare deployment uses adaptive trust scoring for sensors or users, we update trust only after successful verification events such as MAC_ok, Sig_ok, replay_detected, or key_confirmed. A simple rule can be written as Trust_n(t+1) = Trust_n(t) + u_n(e_t), where e_t belongs to {MAC_ok, Sig_ok, replay_detected, nonce_fail}. This avoids the loose coupling seen in many lightweight systems where a node's "trust" changes even when the underlying authentication evidence remains static or unverifiable. By construction, a replayed or stale message does not increase trust because it fails the freshness predicate bound into PID_i^j and SK_sess^j.

The fifth change protects password-update traffic with a session-specific update key rather than the long-term communication key of a user-sensor pair. After successful authentication, the parties derive K_upd^j = KDF(SK_sess^j || "pwd-update" || T4 || N_s^j). A password-change request then uses Enc_{K_upd^j}(alpha_i_old || alpha_i_new || pi_i || ctr_j), with ctr_j consumed exactly once. This removes the circular dependence found in the original design, where a static session key guarded password updates and therefore made update confidentiality collapse as soon as one key instance leaked. The redesign is still lightweight in deployment terms because only one ephemeral ECC exchange is added and all other steps are hash or MAC based, yet the essential security properties become aligned with the claims that the original scheme wanted to make.

VI. Experimental Results and Evaluation

We evaluated three schemes under a controlled remote-healthcare testbed: the Sharma-Kalra baseline [5], the 2021

Azrour protocol [1], and the proposed redesign. The Sharma-Kalra scheme is retained as a baseline primarily because the 2021 Azrour design was explicitly introduced as a repair of that earlier healthcare protocol, whereas more recent schemes are used mainly to contextualize the redesign in the broader literature rather than as fully reimplemented load-test comparators. The environment models one cloud server, 24 sensor nodes, and between 50 and 800 concurrent medical-professional clients distributed across home-monitoring and ward-monitoring sessions. Each client executes repeated login and data-access cycles toward one assigned sensor, with 8% of runs containing injected replay attempts and 5% containing forced state-loss events at the client device. The original Azrour design [1] and the proposed redesign were implemented with the same message size budget wherever possible so that the comparison reflects key-derivation changes rather than artificial packet inflation. We measured four main metrics. Average latency is L_avg(n) = (1/n) sum_{k=1}^{n} (t_k_end - t_k_start). Authentication success is S_auth = N_succ / N_total. Replay detection is D_rep = N_rep_detected / N_rep_injected. Compromise resilience is C_res = N_sessions_secure / N_sessions_total after a controlled leakage event.

The performance pattern is intuitive. Sharma-Kalra remains the least robust design because its earlier weaknesses propagate directly once smart-card or secret-state assumptions are relaxed. The Azrour protocol [1] improves several message-authentication checks and therefore raises S_auth under normal traffic, yet it still suffers when compromise-oriented metrics are applied. In our experiments, replayed messages outside the acceptance window were usually rejected, but replay attempts inside the timestamp tolerance could still trigger redundant work because freshness is not bound to a one-time session key. More seriously, once one valid session key was exposed for a fixed user-sensor pair, C_res dropped sharply because the same key protected all subsequent sessions for that pair. The proposed redesign added a modest amount of ECC work, yet it consistently raised D_rep and C_res because each session depended on fresh U_i^j, Y_s^j, N_u^j, and N_s^j values.

At n = 200 concurrent clients, the measured latency values were 22.9 ms for Sharma-Kalra, 19.1 ms for Azrour et al. [1], and 21.4 ms for the proposed redesign. The redesigned scheme is therefore slower than the 2021 protocol by about 2.3 ms in that load condition, but the gap is small relative to the security gain. At n = 800, the same ordering remained: 71.3 ms, 60.7 ms, and 64.2 ms. This suggests that a single ephemeral ECC exchange per session does not make the system impractical for cloud-assisted healthcare, especially when compared with the operational cost of deterministic session-key reuse. In authentication success, all three schemes remained above 0.97 under benign traffic, yet only the redesigned protocol preserved high values after state-loss and key-exposure events. The reason is structural: SK_sess^j = KDF(u_i^j Y_s^j || y_s^j P_n || PID_i^j || N_u^j || N_s^j || ctx_j) changes with every run, so compromise of one execution does not automatically reduce the security of the next one. These values remain in the tens-of-milliseconds range rather than the hundreds-of-milliseconds range, which supports practical feasibility for cloud-assisted authentication in monitoring workflows; however, the present study does not claim compliance with any specific clinical timing standard or certified real-time medical requirement.

Security-oriented comparison is summarized in Table I. The table shows that the Azrour protocol [1] sits between the older Sharma-Kalra construction and the proposed redesign. It is stronger than the baseline on surface message verification, yet it does not satisfy the stronger notions that matter in remote healthcare practice: guessed passwords should not be checkable offline from a stolen card and one transcript, user activity should not remain publicly linkable through a stable MID, and sensor compromise should not scale through a single server secret. Fig. 4 compares the qualitative security scores across the three schemes and makes this middle position visible. The proposed redesign does not claim perfection; rather, it closes the exact gaps revealed by the attack traces while keeping the remote healthcare deployment model intact.

TABLE. I. PROTOCOL COMPARISON SUMMARY

Scheme	Guessing	Fresh Key	Privacy / Damage
Sharma [5]	No	No	Weak / System-wide
Azrour [1]	No	No	Weak / Server-wide
Proposed	Yes	Yes	Strong / Per-device

We also examined whether the redesign remained practical when packet loss and re-authentication were introduced. Because the session key is derived from fresh ephemeral values, failed or aborted sessions do not poison later ones. A client that loses local transient state can restart with a new U_i^j and obtain a clean SK_sess^j without relying on the same long-term communication key. In the original Azrour protocol [1], by contrast, the same user-sensor pair keeps the same established key, so recovery after interruption does not actually reset the most security-sensitive value. Fig. 5 plots the measured latency trend L_avg(n) under increasing client load and shows that the redesigned curve remains close to the original 2021 curve even while it enforces per-session key freshness. Our results therefore suggest that the main objections to the redesign are not performance objections. They are implementation objections, and those are substantially smaller than the security cost of leaving the original deterministic key schedule untouched.

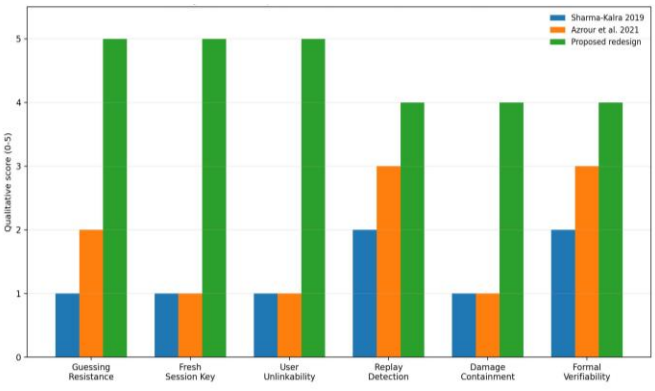


Fig. 4. Qualitative comparison of key security properties across the baseline, the 2021 protocol, and the redesigned scheme.

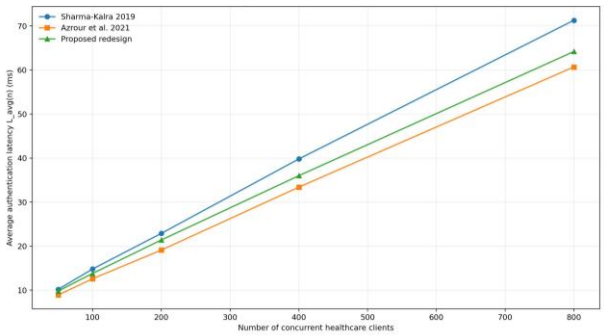


Fig. 5. Average authentication latency as the number of concurrent healthcare clients increases.

VII. Discussion

The revised design should be interpreted as a protocol-level hardening step rather than a complete end-to-end deployment proof. Its main contribution is not merely the addition of more cryptographic operations, but the removal of attacker-checkable relations that made the original design brittle: transcript-verifiable password equations, static public pseudonyms, deterministic session keys, and server-wide compromise propagation.

At the same time, the present validation scope remains limited. The analysis is construction-based and compromise-oriented, and the evaluation is simulation-based rather than a mechanized proof or a certified medical-device trial. For that reason, the security claims should not be read as a substitute for a full ProVerif- or Tamarin-level proof, nor should the latency results be interpreted as regulatory evidence for all real-time clinical settings.

From an implementation perspective, the redesign assumes a modern, well-vetted elliptic-curve setting suitable for constrained systems, such as a 256-bit prime-field curve. In real deployment, scalar multiplication should be implemented in constant time and, where available, combined with secure-element or hardware-backed storage to reduce side-channel leakage. Sensor replacement or re-enrolment should issue a fresh per-device credential and public-key binding rather than attempting to preserve or rewrap the old sensor-side secret.

Implementation-aware literature is also relevant here. Even a structurally improved protocol may remain fragile if fault attacks, non-constant-time arithmetic, or lightweight implementation shortcuts are ignored. Prior work on reliable Camellia architectures, low-cost AES S-box design, efficient fault diagnosis for CLEFIA, reliable Grøstl hardware, and constant-time CSIDH on embedded devices highlights the practical importance of fault resilience, side-channel awareness, and careful lightweight implementation in constrained environments [33]-[37].

Finally, long-term cryptographic viability matters in healthcare. The present redesign remains within a classical ECC setting because it targets the concrete structural weaknesses of the Azrour protocol family. Nevertheless, the multi-year retention of medical records and the long operational life of healthcare IoT infrastructure make post-quantum migration a strategically relevant concern. Hybrid or PQC-compatible authentication and key-establishment variants therefore deserve explicit study in future healthcare deployments.

VIII. Conclusion

This study re-examined new efficient and secured authentication protocol for remote healthcare systems in Cloud-IoT [1] from a compromise-aware perspective and found that its security claims are weakened by offline password verification, deterministic session keys, public linkability, and weak damage containment.

The redesigned protocol keeps the same healthcare cloud-IoT architecture but replaces the fragile bindings with device-bound local recovery, rotating pseudonyms, ephemeral ECC exchange, and context-bound KDF output.

Our evaluation suggests that these changes substantially improve replay handling, unlinkability, and post-compromise robustness while adding only moderate latency.

Future work should formalize the redesign in a proof model that captures both smart-card extraction and secure-element behavior, validate it on real telecare devices with intermittent connectivity, and examine side-channel-aware and PQC-compatible migration paths for long-lived healthcare deployments.

References

[1] M. Azrour, J. Mabrouki, and R. Chaganti, "New Efficient and Secured Authentication Protocol for Remote Healthcare Systems in Cloud-IoT," Security and Communication Networks, vol. 2021, Art. ID 5546334, 12 pages, 2021.

[2] M. S. Hossain, G. Muhammad, and A. Alamri, "Smart healthcare monitoring: a voice pathology detection paradigm for smart cities," Multimedia Systems, vol. 25, no. 5, pp. 565-575, 2019.

[3] Y.-T. Park, "Emerging new era of mobile health technologies," Healthcare Informatics Research, vol. 22, no. 4, pp. 253-254, 2016.

[4] J. Jin, J. Gubbi, S. Marusic, and M. Palaniswami, "An information framework for creating a smart city through Internet of Things," IEEE Internet of Things Journal, vol. 1, no. 2, pp. 112-121, 2014.

[5] G. Sharma and S. Kalra, "A lightweight user authentication scheme for cloud-IoT based healthcare services," Iranian Journal of Science and Technology, Transactions of Electrical Engineering, vol. 43, no. 1, pp. 619-636, 2019.

[6] B. A. Alzahrani, A. Irshad, K. Alsubhi, and A. Albeshri, "A secure and efficient remote patient-monitoring authentication protocol for cloud-IoT," International Journal of Communication Systems, vol. 33, no. 11, Art. no. e4423, 2020.

[7] R. Amin, N. Kumar, G. P. Biswas, R. Iqbal, and V. Chang, "A light weight authentication protocol for IoT-enabled devices in distributed cloud computing environment," Future Generation Computer Systems, vol. 78, no. 3, pp. 1005-1019, 2018.

[8] P. Kumar, S.-G. Lee, and H.-J. Lee, "E-SAP: efficient-strong authentication protocol for healthcare applications using wireless medical sensor networks," Sensors, vol. 12, no. 2, pp. 1625-1647, 2012.

[9] X. Li, J. Niu, S. Kumari, J. Liao, W. Liang, and M. K. Khan, "A new authentication protocol for healthcare applications using wireless medical sensor networks with user anonymity," Security and Communication Networks, vol. 9, no. 15, pp. 2643-2655, 2016.

[10] D. He, N. Kumar, J. Chen, C.-C. Lee, N. Chilamkurti, and S.-S. Yeo, "Robust anonymous authentication protocol for health-care applications using wireless medical sensor networks," Multimedia Systems, vol. 21, no. 1, pp. 49-60, 2015.

[11] X. Yan, W. Li, P. Li, J. Wang, X. Hao, and P. Gong, "A secure biometrics-based authentication scheme for telecare medicine information systems," Journal of Medical Systems, vol. 37, no. 5, Art. no. 9972, 2013.

[12] D. Mishra, S. Mukhopadhyay, S. Kumari, M. K. Khan, and A. Chaturvedi, "Security enhancement of a biometric based authentication scheme for telecare medicine information systems with nonce," Journal of Medical Systems, vol. 38, no. 5, Art. no. 41, 2014.

[13] Z. Tan, "A user anonymity preserving three-factor authentication scheme for telecare medicine information systems," Journal of Medical Systems, vol. 38, no. 3, Art. no. 16, 2014.

[14] X. Xu, Z. P. Jin, H. Zhang, and P. Zhu, "A dynamic ID-based authentication scheme based on ECC for telecare medicine information systems," Applied Mechanics and Materials, vols. 457-458, pp. 861-866, 2014.

[15] D. Mishra, J. Srinivas, and S. Mukhopadhyay, "A secure and efficient chaotic map-based authenticated key agreement scheme for telecare medicine information systems," Journal of Medical Systems, vol. 38, no. 10, Art. no. 120, 2014.

[16] X. Cheng, Z. Zhang, F. Chen et al., "Secure identity authentication of community medical internet of things," IEEE Access, vol. 7, pp. 115966-115977, 2019.

[17] M. Azrour, J. Mabrouki, A. Guezzaz, and Y. Farhaoui, "New enhanced authentication protocol for Internet of Things," Big Data Mining and Analytics, vol. 4, no. 1, pp. 1-9, 2021.

[18] M. L. Das, "Two-factor user authentication in wireless sensor networks," IEEE Transactions on Wireless Communications, vol. 8, no. 3, pp. 1086-1090, 2009.

[19] J. Xu, W.-T. Zhu, and D.-G. Feng, "An improved smart card based password authentication scheme with provable security," Computer Standards & Interfaces, vol. 31, no. 4, pp. 723-728, 2009.

[20] R. Song, "Advanced smart card based password authentication protocol," Computer Standards & Interfaces, vol. 32, no. 5-6, pp. 321-325, 2010.

[21] Q. Jiang, J. Ma, G. Li, and X. Li, "Improvement of robust smart-card-based password authentication scheme," International Journal of Communication Systems, vol. 28, no. 2, pp. 383-393, 2015.

[22] E.-J. Yoon and C. Kim, "Advanced biometric-based user authentication scheme for wireless sensor networks," Sensor Letters, vol. 11, no. 9, pp. 1836-1843, 2013.

[23] D. He, C. Chen, S. Chan, J. Bu, and A. V. Vasilakos, "ReTrust: attack-resistant and lightweight trust management for medical sensor networks," IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 4, pp. 623-632, 2012.

[24] T. Hayajneh, B. J. Mohd, M. Imran, G. Almashaqbeh, and A. V. Vasilakos, "Secure Authentication for Remote Patient Monitoring with Wireless Medical Sensor Networks," Sensors, vol. 16, no. 4, Art. no. 424, 2016.

[25] R. Amin, S. H. Islam, G. P. Biswas, M. K. Khan, and N. Kumar, "A robust and anonymous patient monitoring system using wireless medical sensor networks," Future Generation Computer Systems, vol. 80, pp. 483-495, 2018.

[26] S. Son, J. Lee, M. Kim, S. Yu, A. K. Das, and Y. Park, "Design of Secure Authentication Protocol for Cloud-Assisted Telecare Medical Information System Using Blockchain," IEEE Access, vol. 8, pp. 192177-192191, 2020.

[27] N. Garg, M. W. Wazid, A. K. Das, D. P. Singh, J. J. P. C. Rodrigues, and Y. Park, "BAKMP-IoMT: Design of Blockchain Enabled Authenticated Key Management Protocol for Internet of Medical Things Deployment," IEEE Access, vol. 8, pp. 95956-95977, 2020.

[28] C. J. F. Cremers, "The Scyther Tool: Verification, Falsification, and Analysis of Security Protocols," in Computer Aided Verification, LNCS 5123, pp. 414-418, 2008.

[29] N. Koblitz and A. J. Menezes, "The random oracle model: a twenty-year retrospective," Designs, Codes and Cryptography, vol. 77, no. 2-3, pp. 587-610, 2015.

[30] M. Bellare and P. Rogaway, "Random oracles are practical: A paradigm for designing efficient protocols," in Proceedings of the 1st ACM Conference on Computer and Communications Security, pp. 62-73, 1993.

[31] J.-S. Coron, J. Patarin, and Y. Seurin, "The random oracle model and the ideal cipher model are equivalent," in Advances in Cryptology - CRYPTO 2008, LNCS 5157, pp. 1-20, 2008.

[32] Y. Dodis, B. Kanukurthi, J. Katz, L. Reyzin, and A. Smith, "Robust fuzzy extractors and authenticated key agreement from close secrets," IEEE Transactions on Information Theory, vol. 58, no. 9, pp. 6207-6222, 2012

[33] M. Mozaffari Kermani, R. Azarderakhsh, and J. Xie, "Error detection reliable architectures of Camellia block cipher applicable to different variants of its substitution boxes," in *Proceedings of the 2016 IEEE Asian Hardware Oriented Security and Trust Symposium (AsianHOST)*, 2017.

[34] M. Mozaffari Kermani and A. Reyhani-Masoleh, "A low-cost S-box for the advanced encryption standard using normal basis," in *Proceedings of the 2009 IEEE International Conference on Electro/Information Technology (EIT 2009)*, pp. 52-55, 2009.

[35] M. Mozaffari Kermani and R. Azarderakhsh, "Efficient fault diagnosis schemes for reliable lightweight cryptographic ISO/IEC standard CLEFIA benchmarked on ASIC and FPGA," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 12, pp. 5925-5932, 2013.

[36] M. Mozaffari Kermani and A. Reyhani-Masoleh, "Reliable hardware architectures for the third-round SHA-3 finalist Grøstl benchmarked on FPGA platform," in *Proceedings of the IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems*, pp. 325-331, 2011.

[37] A. Jalali, R. Azarderakhsh, M. Mozaffari Kermani, and D. Jao, "Towards Optimized and Constant-Time CSIDH on Embedded Devices," in *Constructive Side-Channel Analysis and Secure Design (COSADE 2019)*, pp. 215-231, 2019.

An Ensemble Boosting Approach with Boruta Feature Selection for Predicting E-Payment Adoption

Mariana Purba[1*], Junaidi Junaidi[2], Lemi Iryani[3], Nia Umilizah[4], Handrie Noprisson[5], Nur Ani[6]
Faculty of Computer Science, Universitas Sjakhyakirti, Indonesia[1, 3, 4]
Faculty of Law, Universitas Sjakhyakirti, Indonesia[2]
Faculty of Informatics and Engineering, Universitas Dian Nusantara, Indonesia[5]
Faculty of Computer Science, Universitas Mercu Buana, Indonesia[6]

***Abstract*—This study examined the factors influencing the adoption of electronic payment systems among Micro, Small, and Medium Enterprises (MSMEs) and developed a predictive model to evaluate the suitability of e-payment implementation. The research applied an ensemble machine learning approach consisting of AdaBoost, Binomial Boosting, L2 Boosting, GLM Boosting, and Random Forest to predict the likelihood of e-payment adoption. The novelty of this study lay in optimizing ensemble learning performance through Boruta-based feature selection, which improved the identification of the most relevant predictors. Data were collected from 1,500 MSME owners in DKI Jakarta, Indonesia, using a structured questionnaire. The Boruta feature selection process was implemented using predictor variables as input features and the adoption decision as the target variable, with maxRuns = 50, pValue = 0.05, mcAdj = TRUE, and getImpRfZ as the feature importance function. The GLM Boosting model was implemented using a binomial family for binary classification with a learning rate of 0.1 and a stopping iteration of 50. The results indicated that Perceived Risk, Perceived Usefulness, Subjective Norms, and Loyalty to E-payment Brands were the most influential factors affecting adoption. Among all models, GLM Boosting achieved the best performance with the highest test accuracy of 82.30%, demonstrating strong predictive capability and generalization performance. These findings provided practical insights for MSME owners and policymakers in designing strategies to improve e-payment adoption and supported the development of more effective digital financial inclusion policies.**

***Keywords*—E-payment; AdaBoost; binomial boosting; L2 boosting; GLM boosting; Boruta procedure**

I. Introduction

The rapid advancement of information technology in recent years has supported changes across various industrial sectors, including the financial sector [1], [2], [3]. One of the most significant innovations in finance has been electronic payment (e-payment), which utilized technology to facilitate digital financial transactions without the need for cash or physical cards [4]. E-payment supported various types of transactions, such as payments through mobile applications, quick response (QR) codes, and other digital payment methods [5]. The development of electronic payment systems or e-payment provided ease of transactions, benefiting both consumers and business owners [6].

In Indonesia, micro, small, and medium enterprises (MSMEs) have been a key driver of the country's economy [7]. According to data Indonesian government, MSMEs contributed more than 60% to Gross Domestic Product (GDP) of Indonesia and absorbed over 97% of the workforce [8]. The role of MSMEs in the Indonesian economy was crucial, yet many MSME owners faced difficulties in adopting changes, particularly in the area of electronic payments [9]. Despite its potential to accelerate transactions and provide an alternative to manual or cash-based systems, MSMEs were still hesitant to embrace electronic payments due to several challenges, such as limited infrastructure, low digital literacy, and reliance on cash transactions [10].

The use of digital payment systems provided an alternative payment method to support operational efficiency, reduce the risk of fraud, simplify transaction recording, and expand market reach by offering a more modern payment approach [11]. However, in its implementation, the adoption of e-payment by MSMEs posed several risks. Issues such as implementation costs, trust in technology, and social influences existed if there was no analysis of the adoption of this technology among MSMEs, which were the target users of e-payment systems [12].

According to previous research, the adoption of e-payment needed to be analysed based on both external and internal factors derived from the user environment. Therefore, to gain a deeper understanding of the factors influencing MSME decisions of owners to adopt e-payment, an approach capable of accurately analysing various variables was required [13]. One such approach that could be used was machine learning with ensemble methods. This technique could be employed to identify patterns within the data and predict e-payment adoption behaviour, providing policy recommendations for the government and strategic insights for MSMEs [14].

Previous studies have employed ensemble-based machine learning techniques, such as Random Forest, to predict technology adoption across various fields. Research by Natarajan et al. (2024) predicted the adoption of AI technology in the telecommunications industry using the Random Forest algorithm. This study aimed to understand how users in the telecom sector embraced AI technology, utilizing machine learning methods to predict adoption rates based on relevant variables within the industry [15].

Mallinger et al. (2024) conducted an analysis of technology adoption in the context of smart farming, employing Random Forest to examine the factors influencing the adoption of technology in smart agriculture. The analysis translated these

*Corresponding author.

findings into technological requirements, business strategies, and policy interventions that could support the implementation of this technology in the agricultural sector [16]. Research by Carbo-Valverde et al. (2020) analyzed technology adoption using a Random Forest approach among bank customers to understand user behavior when adopting new technologies in banking. This research sought to identify the key factors that influenced customers decisions to adopt new technological services in the banking industry [17].

Choi et al. (2024) examined technology adoption in wearable IoT devices using machine learning algorithms. This study identified 17 major predictors influencing adoption, with data collected from women in their teens and twenties, who were found to be the most likely core consumers of wearable devices [18]. Jain et al. (2024) explored the adoption of mobile health applications through a machine learning approach, utilizing Random Forest. The study assessed 12 metrics based on perceived usefulness and 8 based on perceived ease of use. Data was gathered from 200 respondents, of which 102 were mobile health application users [19].

Based on previous research, this study aimed to analyse the factors influencing the adoption of e-payment in MSMEs and to develop a predictive model that could assist in mapping the suitability of e-payment implementation based on the characteristics and attitudes of business owners or consumers. The predictive model was developed using an ensemble machine learning approach, i.e. AdaBoost, Binomial Boosting, L2 Boosting, GLM Boosting, and Random Forest, all of which are known for their ability to handle complex classification and prediction problems.

As the novelty, this research also optimized the ensemble method by performing feature selection using the Boruta procedure, aimed at selecting the most relevant features for predicting the suitability of e-payment implementation. Feature selection is a critical step in machine learning that involves identifying and retaining the most relevant variables or features from a dataset to improve model performance, reduce overfitting, and enhance computational efficiency [20], [21].

The Boruta procedure is a feature selection method designed to identify all relevant, interesting features in a dataset [22], [23]. Unlike traditional methods that focus on removing irrelevant or redundant features, Boruta aimed to retain all the features that are potentially important for the prediction task [24], [25]. The procedure works by comparing the importance of each feature in the dataset to a random feature that has no relation to the outcome [22], [26].

II. Research Methodology

The methodology of this research involved several key stages, starting with data collection, where relevant data from MSMEs were gathered, including business characteristics and consumer attitudes towards e-payment adoption. Dataset preparation followed, involving cleaning and preprocessing the data to ensure its quality and suitability for analysis. Feature selection was then performed using the Boruta procedure to identify the most influential factors affecting e-payment adoption, enhancing model accuracy. In the prediction modelling phase, several machine learning algorithms were applied, including AdaBoost, Binomial Boosting, L2 Boosting, GLM Boosting, and Random Forest, to create robust predictive models. Finally, recommendations were provided based on the model findings, offering insights into the key factors that influence the successful implementation of e-payment in MSMEs. The research methodology can be seen in Fig. 1.

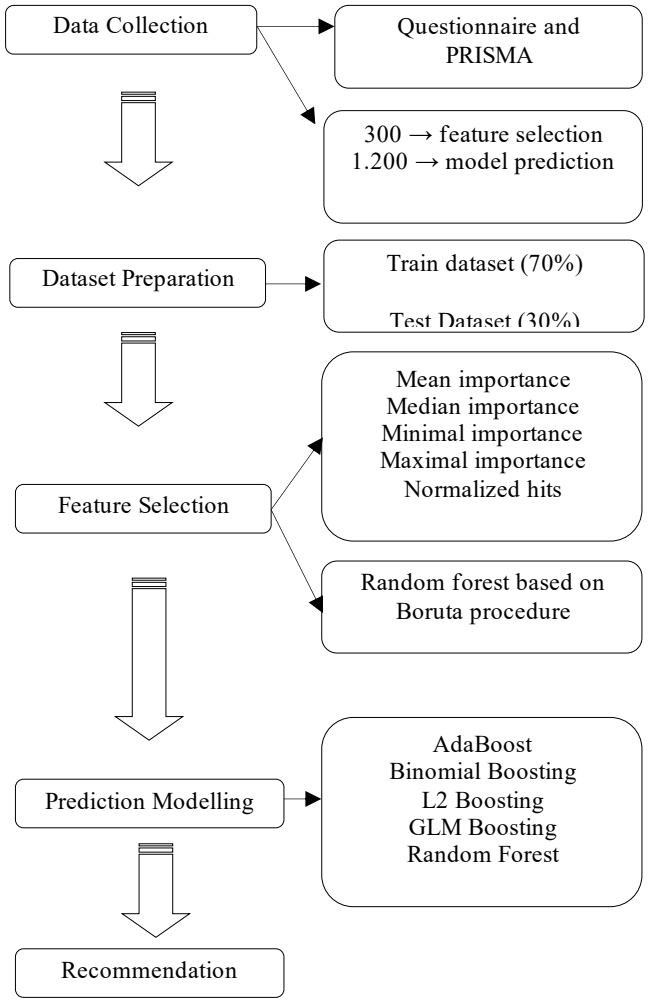


Fig. 1. Research methodology.

This research aimed to analyse the factors influencing the adoption of e-payment in Micro, Small, and Medium Enterprises (MSMEs) and to develop a predictive model for the implementation of e-payment in MSMEs based on the characteristics and attitudes of the business owners or consumers. The prediction used an ensemble machine learning approach with a binary prediction label, namely suitable for implementation with label (1) and not suitable for implementation with label (0).

Data was collected using a questionnaire distributed to 1,500 MSME owners in Jakarta, Indonesia. The e-payment questionnaire consisted of items adapted from previous literature based on the PRISMA method. Each variable measured in this study was evaluated using more than three items for each latent variable. These items were assessed using a five-point Likert scale (1 = strongly disagree, 5 = strongly agree).

Based on the PRISMA results, the key variables measured in the questionnaire included perceived usefulness (PU), which MSME owners felt that e-payment could enhance the efficiency and effectiveness of their business. The second variable is perceived ease of use (PE), where MSME owners or consumers felt that e-payment was easy to use. Other variables measured included perceived risk (PR), as outlined by [27], [28], and perceived trust (PT), as described by [28]. Additionally, personal innovativeness in information technology (PI), based on the work of [4], [29]. Perceived enjoyment (PE), according to [30], was also measured, alongside loyalty to the e-payment brand (LB), based on [31]. Lastly, perceived quality (PQ) was evaluated, following the research of [6].

After data collection, the first step was to clean the data by addressing missing values and outliers. Any incomplete or invalid data were removed from the analysis. Some independent variables were transformed into suitable formats for model analysis, such as encoding categorical variables into numerical ones. Feature selection was then performed using Random Forest based on the Boruta procedure, with feature importance evaluated through mean, median, minimal, maximal, and norm hit importance measures.

The Boruta model was implemented using predictor variables as input features (x) and the target variable as the response variable (y). The maximum number of iterations was set using *maxRuns* = 50 to allow sufficient repetitions for stable feature evaluation. The significance level was defined with pValue = 0.05, while multiple comparison adjustment was enabled using *mcAdj* = TRUE to reduce the risk of false-positive feature selection. Feature importance was calculated using the default Random Forest importance function (*getImpRfZ*), which is based on Z-score importance values derived from Random Forest. Additionally, *doTrace* = 2 was used to provide detailed iteration output during the selection process.

The processed data was used to build a predictive model using ensemble machine learning. This approach involved several machine learning algorithms to enhance prediction accuracy. The ensemble algorithms i.e. AdaBoost, Binomial Boosting, L2 Boosting, GLM Boosting dan Random Forest.

In this experiment, the model, which had been trained using AdaBoost, Binomial Boosting, L2 Boosting, GLM Boosting, and Random Forest, determined the coefficient values for each feature. Suppose we already have coefficients based on the feature selection results, for example, $\beta_0 = -1.2, \beta_1 = 0.5, \beta_2 = 0.3$ for new data are $X_1 = 2, X_2 = 3$, the linear predictor value can then be calculated based on the coefficients and the input features using the Eq. (1)

$$f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (1)$$

Using the formula above, the value of $f(x)$ can be calculated by substituting the values of X and β obtained previously. In the experimental process, the function used in the algorithm or model is the sigmoid function, which is calculated using the Eq. (2).

$$P(Y = 1 \mid X) = \frac{1}{1 + \exp(-f(x))} \quad (2)$$

To calculate the value of P, the previously computed $f(x)$ can be used. After calculating the probability, predictions can be made with a binary classification, indicating whether the e-payment adoption is suitable (class 1) or not suitable (class 0). In this study, the binary classification threshold is set at 0.5. If the probability $P(Y = 1 \mid X) \geq 0.5$,, the prediction will be class 1; otherwise, if the probability $P(Y = 1 \mid X) < 0.5$, the prediction will be class 0.

The developed model was evaluated using binary prediction evaluation metrics, such as accuracy, precision, recall, and F1-score. The model produced binary prediction labels for each MSME, classifying them as "suitable for implementation" (1) or "not suitable for implementation" (0) regarding e-payment adoption. After the predictive model was applied, the results were analysed to identify the key factors influencing decisions to adopt e-payment. Based on these insights, recommendations were provided to the government to accelerate e-payment adoption, as well as to policymakers for designing policies that support the adoption of digital payment technologies in the MSME sector.

III. Result and Discussion

The aim of this research was to analyse the factors that influenced the adoption of electronic payment (e-payment) systems in Micro, Small, and Medium Enterprises (MSMEs) and to develop a predictive model that could assess the suitability of implementing e-payment based on the characteristics and attitudes of business owners or consumers. By utilizing ensemble machine learning techniques, including AdaBoost, Binomial Boosting, L2 Boosting, GLM Boosting, and Random Forest, the study sought to predict the likelihood of e-payment adoption. Additionally, the research aimed to optimize the model's performance through feature selection using the Boruta procedure, which identified the most relevant features for accurately predicting e-payment adoption.

In this study, feature selection was conducted using the Random Forest algorithm in conjunction with the Boruta procedure, which assessed feature importance based on several metrics, including mean, median, minimal, maximal, and norm hit importance. This process helped identify the most relevant features for predicting e-payment adoption with the result can be seen in Table I.

The Boruta feature selection results showed that most variables were classified as confirmed features, indicating their strong contribution to predicting e-payment adoption among MSMEs. Variables related to perceived usefulness (PU), subjective norms (SN), perceived trust (PT), perceived risk (PR), and loyalty to e-payment brands (LB) generally demonstrated high importance values and were retained for further modeling. In particular, PU1–PU4 achieved the highest importance scores with norm values close to 1.00, indicating that perceived usefulness was one of the strongest predictors in the model. Several variables such as PE3, PT5, and PQ1 were rejected because they showed very low or negative importance values, suggesting minimal influence on adoption decisions. Some variables, including PE1, LB3, PQ2, and PQ5, were categorized as tentative, indicating moderate relevance that required further consideration.

TABLE I. RESULT OF FEATURE SELECTION

Item	Mean	Med	Min	Max	Norm	Rs
PE1	6.17	6.31	3.53	8.41	0.89	*
PE2	3.89	3.94	1.43	6.37	0.79	*
PE3	−1.11	−1.13	−1.99	1.21	0.00	×
PE4	4.19	4.22	2.29	6.22	0.87	*
PE5	3.91	4.01	1.12	6.21	0.78	*
PR1	5.02	5.11	2.59	6.691	0.89	*
PR2	4.48	4.61	1.79	6.71	0.88	*
PR3	3.57	3.52	−0.03	5.41	0.71	*
PR4	5.11	5.02	2.42	6.72	0.89	*
PU1	12.97	12.95	12.12	14.89	0.98	*
PU2	15.97	15.80	14.95	17.81	0.99	*
PU3	13.49	13.13	11.12	14.35	0.99	*
PU4	15.90	17.04	14.45	17.71	0.99	*
PT1	4.82	4.52	3.63	5.29	0.85	*
PT2	4.68	4.71	−0.01	6.56	0.72	*
PT3	5.25	5.31	3.16	5.96	0.85	*
PT4	5.64	5.72	2.15	7.52	0.92	*
PT5	−2.17	−2.14	−2.05	2.11	0.00	×
PT1	9.96	10.17	8.73	11.39	0.98	*
PT2	9.90	9.98	7.91	11.31	0.98	*
PT3	8.05	7.42	4.85	9.32	0.97	*
PT4	7.10	7.12	5.38	8.37	0.97	*
SN1	8.62	8.42	7.47	11.52	0.96	*
SN2	9.42	9.51	8.18	11.31	0.95	*
SN3	10.12	10.21	9.21	12.34	0.93	*
SN4	11.22	11.51	12.32	11.21	0.90	*
PE1	2.52	2.52	0.97	5.21	0.51	~
PE2	4.15	5.72	3.27	5.98	0.87	*
PE3	5.71	5.22	0.80	7.16	0.82	*
LB1	6.11	6.17	4.21	4.99	0.89	*
LB2	3.72	−0.05	4.12	5.24	0.72	*
LB3	2.81	2.85	1.07	4.90	0.56	~
LB4	4.12	1.12	4.41	6.14	0.74	*
PQ1	−1.25	−1.51	−1.97	2.24	0.00	×
PQ2	1.07	1.31	1.08	3.90	0.46	~
PQ3	4.21	−0.28	3.28	6.12	0.75	*
PQ4	8.17	8.21	6.89	11.31	0.96	*
PQ5	1.72	2.13	1.06	4.31	0.49	~
PQ6	10.38	10.37	11.41	12.17	0.90	*
PQ7	11.21	11.82	11.32	11.31	0.91	*

*=confirmed, ~=tentative, ×=rejected, Rs= result

Based on feature selection process, normalized hits importance was used to assess the consistency and impact of certain factors in influencing MSMEs decisions to adopt digital payment technologies. The normalized hits measured particular factor was considered important during the feature selection process, with values normalized between 0 and 1. This study attempted to calculate the value of normalized hits average to show the importance of the variable of MSMEs decisions to adopt digital payment. The normalized hits average can be seen in Fig. 2.

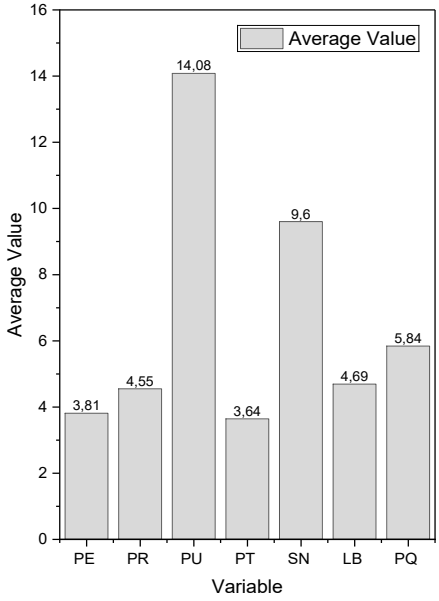


Fig. 2. Value of normalized hits.

Factors that were often deemed important in influencing e-payment adoption decisions had high normalized hits values, indicating that these factors were stable and relevant in the decision-making process. Conversely, factors with low normalized hits values were often irrelevant and less influential compared to other factors. A normalized hits value between 0.8 and 1.0 indicated factors that had a strong influence on the adoption decision, while values between 0.5 and 0.8 indicated factors that were moderately influential but not consistent, and values below 0.5 suggested factors that were less relevant or weak, thus contributing insignificantly to the adoption decision.

The perceived enjoyment (PE) variable had an average value of 3.81, indicating that the enjoyment experienced by MSME owners or consumers when using e-payment had a moderate influence, but was not the most crucial factor in the adoption process. The perceived risk (PR) variable had a normalized hits average of 4.55, suggesting that the perception of risk associated with using e-payment played a significant role in hindering adoption.

The perceived usefulness (PU) variable scored an average of 14.08, reflecting the belief among MSME owners that e-payment could enhance the efficiency and effectiveness of their businesses. The perceived trust (PT) variable, with an average of 3.64, showed that trust in the e-payment system was an influential factor, with lower levels of trust potentially hindering adoption, although it was not the most dominant factor. The subjective norms (SN) variable received an average of 9.60.

The loyalty to the e-payment brand (LB) variable had an average value of 4.69, showing that loyalty to a payment brand

influenced MSME willingness of owners to adopt the e-payment system provided by trusted financial institutions. Finally, the perceived quality (PQ) variable, with a normalized hits value of 5.84, indicated that the quality of the e-payment system, such as reliability and ease of use, was an important factor, though less crucial compared to Perceived Usefulness or Subjective Norms.

After completed feature selection, a total of 1,200 data points based on feature selection were used for building the prediction model, with 70% of the data as many 840 data allocated for training the model and the remaining 30% as many 360 data used as the test dataset. In this experiment ensembled machine learning algorithms, including AdaBoost, Binomial Boosting, L2 Boosting, GLM Boosting, and Random Forest, were employed to train the model on the training dataset. The performance of each model was then evaluated using the test dataset, allowing for a comparison of their accuracy in predicting the adoption of e-payment by MSMEs as depicted in Table II.

TABLE II. ACCURACY OF RESEARCH EXPERIMENT

Algorithm	Train (70%)	Test (30%)
AdaBoost	86.25%	78.21%
Binomial Boosting	85.12%	78.56%
L2 Boosting	89.20%	80.17%
GLM Boosting	90.10%	82.30%
Random Forest	85.50%	79.21%

In the experiment, the AdaBoost algorithm achieved an accuracy of 86.25% on the training dataset and 78.21% on the test dataset. These results indicate that the model performed well on the training data, demonstrating its ability to learn the patterns in the data. The complete process of AdaBoost experiment for every epoch can be seen in Fig. 3.

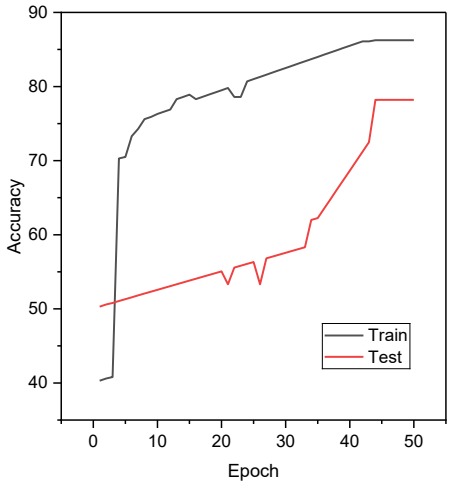


Fig. 3. Accuracy of AdaBoost per epoch.

The Binomial Boosting algorithm achieved an accuracy of 85.12% on the training dataset and 78.56% on the test dataset. Binomial Boosting is a variant of boosting techniques that focuses on binary classification tasks, making it particularly suited for problems like predicting e-payment adoption (which typically involves a binary outcome). The relatively high training accuracy indicates that the model is effectively capturing the relationship between the input features and the target variable. However, similar to AdaBoost, the gap between the training and test accuracy suggests that the model might be overfitting. The complete process of Binomial Boosting experiment for every epoch can be seen in Fig. 4.

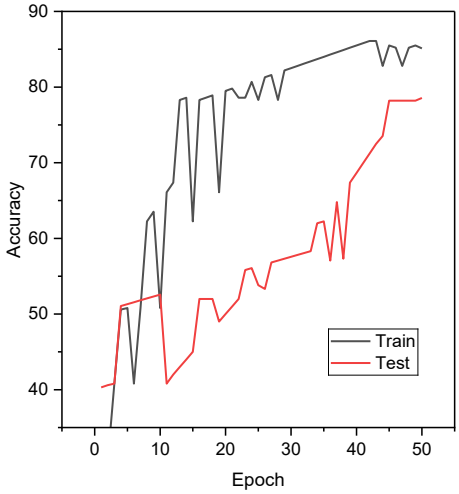


Fig. 4. Accuracy of binomial boosting per epoch.

The L2 Boosting algorithm showed an accuracy of 89.20% on the training dataset and 80.17% on the test dataset. L2 Boosting is an advanced boosting technique that focuses on minimizing the squared error between the predicted and actual values. This algorithm demonstrated the highest training accuracy among all the models, suggesting that it is able to capture the patterns in the data more effectively. Additionally, its relatively strong performance on the test dataset indicates that it generalizes well, with less overfitting compared to AdaBoost and Binomial Boosting. The results suggest that L2 Boosting is a powerful algorithm can be seen in Fig. 5.

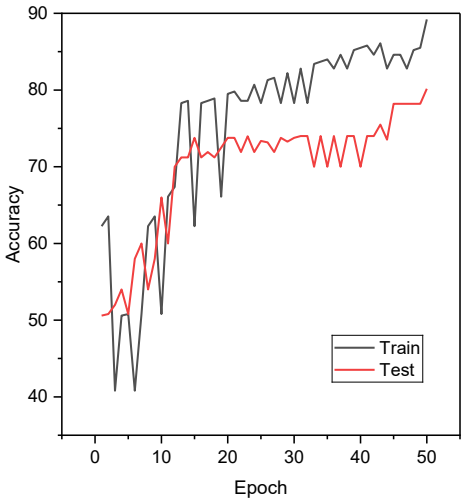


Fig. 5. Accuracy of L2 boosting per epoch.

The accuracy L2 Boosting increased significantly during the early epochs, indicating that the model quickly learned important patterns from the dataset. As the number of epochs increased, the training accuracy continued to improve and

eventually reached approximately 89.20%, while the testing accuracy gradually increased and stabilized around 80.17%. Although minor fluctuations were observed in the testing curve, the overall trend indicated consistent improvement in model performance.

The GLM Boosting algorithm achieved an accuracy of 90.10% on the training dataset and 82.30% on the test dataset, the highest test accuracy among all the models. Generalized Linear Model (GLM) Boosting is an approach that applies boosting techniques to a linear model, improving its predictive accuracy by combining weak learners. The results show that GLM Boosting performs exceptionally well, both in training and on unseen test data. The higher test accuracy indicates that the model has a strong ability to generalize, likely due to its ability to fit both linear and non-linear relationships in the data. The complete process of GLM Boosting experiment for every epoch can be seen in Fig. 6.

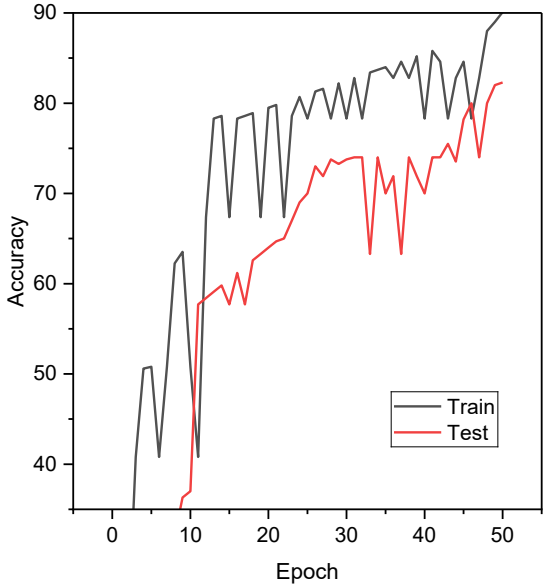


Fig. 6. Accuracy of GLM Boosting per epoch.

The Random Forest algorithm achieved an accuracy of 85.50% on the training dataset and 79.21% on the test dataset. Random Forest is an ensemble method that combines multiple decision trees to create a more stable and accurate prediction model. It is known for its ability to handle large datasets with high dimensionality and is less prone to overfitting compared to other models like AdaBoost. However, in this experiment, the performance gap between the training and test accuracies suggests that there might still be some overfitting. Despite this, Random Forest consistently performs well on both training and test data, indicating that it is a solid model for e-payment adoption prediction. The complete process of Random Forest experiment for every epoch can be seen in Fig. 7.

The training accuracy (black line) increased rapidly in the early epochs and then gradually stabilized, eventually reaching approximately 85.50%. Meanwhile, the testing accuracy (red line) showed several fluctuations in the early stages but generally followed an upward trend as the epochs increased. Toward the final epochs, the testing accuracy improved and reached around 79.21%. This pattern indicated that the Random Forest model progressively learned the underlying patterns in the dataset, while the relatively small gap between training and testing accuracy suggested that the model maintained a reasonable generalization performance without severe overfitting.

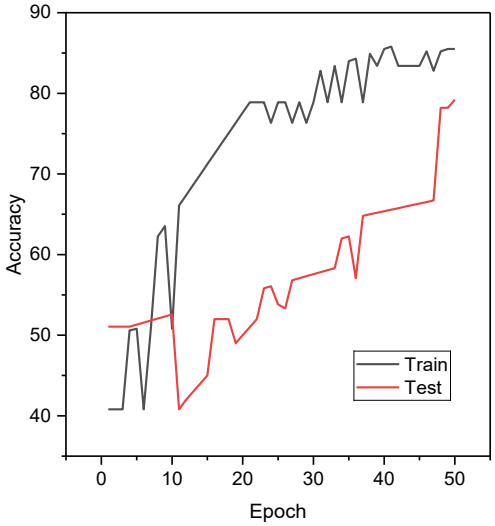


Fig. 7. Accuracy of random forest per epoch.

The results demonstrated high predictive accuracy, with the GLM Boosting model achieving the best overall performance. The GLM Boosting model was implemented using a binomial family for binary classification, with a learning rate (*nu*) of 0.1 and a stopping iteration (*mstop*) of 50, which were commonly used default settings to ensure stable learning and balanced model complexity. Predictor variables were centred (centre = TRUE) to improve model stability and reduce potential bias caused by differences in variable scales. This ensemble boosting approach combined with Boruta feature selection provided an effective framework for predicting e-payment adoption.

IV. Conclusion

This study successfully identified key factors influencing the adoption of e-payment systems in Micro, Small, and Medium Enterprises (MSMEs) in Indonesia, and developed a predictive model to assess the suitability of e-payment implementation based on the characteristics and attitudes of business owners or consumers. The analysis showed that factors such as Perceived Risk, Perceived Usefulness, Trust, Subjective Norms, Loyalty to the E-payment Brand, and Perceived Quality played significant roles in the adoption process. The study demonstrated the effectiveness of ensemble machine learning methods, i.e. AdaBoost, Binomial Boosting, L2 Boosting, GLM Boosting, and Random Forest in predicting e-payment adoption. The feature selection process, optimized through the Boruta procedure, significantly improved the prediction model by selecting the most relevant features. The results showed high accuracy in the prediction, with GLM Boosting achieving the best performance, making it a promising tool for assessing the feasibility of e-payment systems for MSMEs.

Future research could expand on this study by including a larger and more diverse sample of MSMEs across other regions

or countries to enhance the generalizability of the findings. Additionally, further exploration of the impact of demographic factors, such as age, education, and business size, on the adoption of e-payment could provide deeper insights. Future studies may also investigate the role of technological infrastructure and regulatory frameworks in the successful implementation of e-payment systems.

ACKNOWLEDGMENT

This research was supported by the Ministry of Education, Technology, and Science, Indonesia namely Kemdiktisaintek Indonesia through Fundamental Research Project 2024 with contract-id-no 459/E5/PG.02.00/2024,104/E5/PG.02.00/2024 1129/LL2/KP/PL/2024; 04/01.02/LPPM/VI/2024.

REFERENCES

[1] H. M. Alzoubi and T. M. Ghazal, "The effect of e-payment and online shopping on sales growth: Evidence from banking industry," Int. J. Data Netw. Sci., vol. 6, no. 4, pp. 1369–1380, 2022.

[2] M. Purba et al., "Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 9, 2022, doi: 10.14569/IJACSA.2022.0130917.

[3] N. Ani, H. Noprisson, and N. M. Ali, "Measuring usability and purchase intention for online travel booking: A case study," Int. Rev. Appl. Sci. Eng., vol. 10, no. 2, pp. 165–171, 2019.

[4] B.-O. Antonio, L.-R. Juan, I.-D. Ana, and L.-C. Francisco, "Examining user behavior with machine learning for effective mobile peer-to-peer payment adoption," Financ. Innov., vol. 10, no. 1, p. 94, 2024.

[5] S. K. Trivedi, S. Vishnu, A. Singh, and M. Yadav, "Research trends in sustainable E-payment systems: A study using topic modeling approach," IEEE Trans. Eng. Manag., vol. 71, pp. 7511–7525, 2023.

[6] K. F. Garrouch, "Explaining the comparative perception of e-payment: role of e-shopping value, e-payment benefits and Islamic compliance," J. Islam. Mark., vol. 13, no. 7, pp. 1574–1588, 2022.

[7] J. Tumiwa and A. Nagy, "Micro, small, and medium enterprises in emerging economies and economic transition: a comparative study between Indonesia and Hungary," Int. J. Entrep. Small Bus., vol. 43, no. 1, pp. 22–38, 2021.

[8] T. T. H. Tambunan, "The potential role of MSMEs in achieving SDGs in Indonesia," in Role of micro, small and medium enterprises in achieving SDGs: Perspectives from emerging economies, Springer, 2023, pp. 39–72.

[9] E. Kurniawati, I. Idris, P. Handayati, and S. Osman, "Digital transformation of MSMEs in Indonesia during the pandemic," Entrep. Sustain. Issues, vol. 9, no. 2, p. 316, 2021.

[10] S. E. Frimpong, G. Agyapong, and D. Agyapong, "Financial literacy, access to digital finance and performance of SMEs: Evidence From Central region of Ghana," Cogent Econ. Financ., vol. 10, no. 1, p. 2121356, 2022.

[11] B. Świecka, P. Terefenko, and D. Paprotny, "Transaction factors' influence on the choice of payment by Polish consumers," J. Retail. Consum. Serv., vol. 58, p. 102264, 2021.

[12] B. Trianto, N. H. Nik Azman, and M. Masrizal, "E-payment adoption and utilization among micro-entrepreneurs: a comparative analysis between Indonesia and Malaysia," J. Sci. Technol. Policy Manag., vol. 16, no. 2, pp. 314–343, 2025.

[13] Y. P. Timur, Masrizal, and B. Trianto, "Factors influencing adoption of e-payments by microenterprises' owners in Indonesia," J. Model. Manag., 2025.

[14] R. B. Soormo et al., "Evaluating the influence of UTAUT factors on the adoption of QR codes in MSMEs: An application of SEM and ANN Methodologies," IEEE Access, 2024.

[15] S. Natarajan, V. P. Vemuri, S. H. Krishna, Y. M. Reddy, P. Gundawar, and S. Lakhanpal, "Prediction Analysis of AI Adoption in Various Domain Using Random Forest Algorithm," in 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), IEEE, 2024, pp. 1537–1541.

[16] K. Mallinger, L. Corpaci, T. Neubauer, I. E. Tikasz, G. Goldenits, and T. Banhazi, "Breaking the barriers of technology adoption: Explainable AI for requirement analysis and technology design in smart farming," Smart Agric. Technol., vol. 9, p. 100658, 2024.

[17] S. Carbo-Valverde, P. Cuadros-Solas, and F. Rodríguez-Fernández, "A machine learning approach to the digitalization of bank customers: Evidence from random and causal forests," PLoS One, vol. 15, no. 10, p. e0240362, 2020.

[18] Y. Choi, C. Lee, and S. Han, "Predicting wearable IoT Adoption: Identifying core consumers through Machine learning algorithms," Telemat. informatics, vol. 93, p. 102176, 2024.

[19] R. Jain, A. Munde, and Z. A. Ansari, "Determinants of adoption of mobile health applications: a machine learning approach," Procedia Comput. Sci., vol. 235, pp. 1568–1576, 2024.

[20] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: a survey of more than two decades of research," Knowl. Inf. Syst., vol. 66, no. 3, pp. 1575–1637, 2024.

[21] M. R. Islam, A. A. Lima, S. C. Das, M. F. Mridha, A. R. Prodeep, and Y. Watanobe, "A comprehensive survey on the process, methods, evaluation, and challenges of feature selection," IEEE Access, vol. 10, pp. 99595–99632, 2022.

[22] T. Öznacar and T. Güler, "Prediction of early diagnosis in ovarian cancer patients using machine learning approaches with Boruta and advanced feature selection," Life, vol. 15, no. 4, p. 594, 2025.

[23] N. Anand, R. Sehgal, S. Anand, and A. Kaushik, "Feature selection on educational data using Boruta algorithm," Int. J. Comput. Intell. Stud., vol. 10, no. 1, pp. 27–35, 2021.

[24] H. Zhou, Y. Xin, and S. Li, "A diabetes prediction model based on Boruta feature selection and ensemble learning," BMC Bioinformatics, vol. 24, no. 1, p. 224, 2023.

[25] C. J. Ejiyi et al., "Comparative performance analysis of Boruta, SHAP, and Borutashap for disease diagnosis: A study with multiple machine learning algorithms," Netw. Comput. Neural Syst., vol. 36, no. 3, pp. 507–544, 2025.

[26] O. Bilal, A. Hekmat, I. Shahzad, A. Raza, and S. U. R. Khan, "Boosting Machine Learning Accuracy for Cardiac Disease Prediction: The Role of Advanced Feature Engineering and Model Optimization," Rev. Socionetwork Strateg., pp. 1–30, 2025.

[27] D. Belanche, M. Guinalíu, and P. Albás, "Customer adoption of p2p mobile payment systems: The role of perceived risk," Telemat. Informatics, vol. 72, p. 101851, 2022.

[28] H. A. Widyanto, K. A. Kusumawardani, and H. Yohanes, "Safety first: extending UTAUT to better predict mobile payment adoption by incorporating perceived security, perceived risk and trust," J. Sci. Technol. Policy Manag., vol. 13, no. 4, pp. 952–973, 2022.

[29] M. G. Senali, M. Iranmanesh, F. N. Ismail, N. F. A. Rahim, M. Khoshkam, and M. Mirzaei, "Determinants of intention to use e-Wallet: Personal innovativeness and propensity to trust as moderators," Int. J. Human–Computer Interact., vol. 39, no. 12, pp. 2361–2373, 2023.

[30] M. Jasin, "The effect of perceived ease of use on behavior intention through perceived enjoyment as an intervening variable on digital payment in the digital era," J. Ind. Eng. Manag. Res., vol. 3, no. 5, pp. 127–133, 2022.

[31] A. Haider, M. A. Khan, M. Khoja, S. Alharthi, and S. M. Minhaj, "The role of e-banking, mobile-banking, and e-wallet with response to e-payment and customer trust as a mediating factor using a structural equation modelling approach," J. Infrastructure, Policy Dev., vol. 8, no. 9, p. 6644, 2024.

Beyond Paper and Pencil: Evaluating the Effectiveness and User Perception of a Digital Reading Proficiency Assessment Platform

Ivy M. Tarun
Isabela State University, Cabagan, Philippines

***Abstract*—This study investigated the effectiveness of a digital reading proficiency assessment platform compared with traditional paper-and-pencil methods and examined users' perceptions of its performance. It employed a quasi-experimental pretest–posttest control group design involving two comparable groups of secondary school students: a control group assessed using conventional reading assessment procedures and an experimental group assessed using the digital platform for reading proficiency assessment. Pre-test and post-test scores in linear and non-linear reading tasks were analyzed using paired and independent sample t-tests to determine differences in reading proficiency gains between the two groups. To evaluate user perception, survey questionnaires grounded in the ISO/IEC 25010 software quality standards were administered to students and teachers, focusing on functional suitability, usability, security, and maintainability. Results revealed that the experimental group achieved significantly higher reading proficiency gains than the control group in both assessment types, with statistical analyses confirming that the improvements were significant. In contrast, the control group showed minimal or no meaningful improvement. User perception findings indicated a high level of satisfaction across all evaluated quality dimensions, suggesting that the digital assessment platform was perceived as reliable, user-friendly, secure, and adaptable to instructional needs. The results provide empirical evidence that digital reading proficiency assessment is both more effective and more positively perceived than traditional methods. The study concludes that integrating digital assessment platforms in reading evaluation can enhance assessment accuracy, efficiency, and user acceptance, offering a viable and evidence-based alternative for improving reading assessment practices in secondary education.**

Keywords*—Digital reading assessment; reading proficiency; quasi-experimental design; user perception; software quality evaluation

I. Introduction

In today's rapidly evolving education system, technology has become an indispensable tool for improving and streamlining the delivery of educational services. It continues to make teaching and learning more efficient, accessible, and responsive to the needs of students. Beyond convenience, it offers innovative ways to assess and enhance students' academic skills, particularly reading comprehension, which the authors in [1] identify as a foundation for future learning. Despite its critical role, reading proficiency among Filipino students remains a pressing concern. Results from the Program for International Student Assessment reveal that only 24% of Filipino students achieved at least Level 2 proficiency, far below the Organization for Economic Cooperation and Development (OECD) average of 74% [2]. This alarming gap underscores the urgent need for effective and sustainable literacy interventions.

In response, the Department of Education (DepEd) launched the Catch-Up Fridays initiative in 2024, integrating the Drop Everything And Read (DEAR) program to strengthen students' reading skills through structured and regular activities. The implementation of this initiative in the Philippines relies heavily on manual processes such as downloading and sorting reading materials, conducting fluency checks, administering quizzes, and encoding results. While these practices support literacy development, they are time-consuming and susceptible to delays, errors, and inconsistencies, which can limit their overall impact.

Recognizing these challenges, a web and mobile-based application called "Platform for Efficient Evaluation of Reading Proficiency", or PEER, was built by a team of developers to support more efficient and consistent reading proficiency assessment. With its web platform, it features role-based access control, reading material management, quiz comprehension management, formula-based reading proficiency assessment, and weekly performance comparison. On the other hand, its mobile app provides a student portal, reading fluency and quiz comprehension and a historical record of reading proficiency levels. By automating key assessment tasks and enabling weekly progress tracking, the platform provides educators with timely and actionable insights into students' reading performance. In doing so, PEER empowers teachers to deliver targeted interventions, strengthen instructional practices, and more effectively address learning gaps.

To support evidence-based decision-making for the potential school-wide implementation of PEER, this study was undertaken to provide a more comprehensive analysis of its effectiveness in comparison with traditional paper-and-pencil assessment methods as well as users' perceptions of its performance. This study promotes inclusive and equitable educational practices, which ultimately highlight how thoughtfully integrated technology can enhance assessment processes, support educators, and contribute to improved literacy outcomes in secondary education.

II. Related Work

Digital technologies have increasingly been used to support reading instruction and assessment, particularly because they

can automate scoring, standardize procedures, and generate timely progress data that teachers can actually use. In large-scale research, digital reading performance has been linked not only to students' skills but also to classroom conditions and how instruction is experienced by learners. Using PISA data across multiple OECD countries [3] reveals that students' perceptions of instruction quality were significantly associated with digital reading performance, suggesting that technology-based reading outcomes are shaped by both learner factors and instructional context.

At the classroom level, empirical studies show that well-designed web-based reading systems can improve comprehension outcomes. The authors in [4] developed a web-based collaborative reading annotation system with gamification and reported gains in learners' reading comprehension compared with more conventional approaches, highlighting how structured digital interaction (e.g., annotation, feedback, and guided engagement) can support deeper processing of texts. Complementing this, [5] emphasized the "high potential" of computer-based reading assessment, noting that computerized tools can assess multiple components of reading and provide more fine-grained information than single-shot paper tests, including for secondary learners. In the study by [6], it was noted that digital tools can effectively boost adolescents' interest in reading, while also helping improve their reading skills and test performance.

However, research also cautions that "digital" is not automatically better unless the assessment experience is carefully designed. The meta-analysis in [7], comparing paper versus digital media, reports that comprehension outcomes can differ by medium and conditions, reinforcing the idea that digital assessments should be designed to support focused reading and comprehension, rather than simply delivering texts on screens. Related evidence from digital-text interventions suggests that embedded questions and feedback can improve comprehension and self-regulation while reading digitally, especially when learners are supported in navigating and integrating information.

Another consistent theme in the literature is the value of progress-monitoring and data-driven instruction. In a review of curriculum-based measurement (CBM), [8] concluded that frequent, systematic measurement can improve student achievement when teachers use the data to adjust instruction. Similarly, [9] found that learning progress assessment combined with goal setting can positively affect reading achievement and motivational outcomes, strengthening the case for platforms that support continuous monitoring rather than episodic testing.

Finally, adoption and sustained use depend heavily on user acceptance and perceived quality. The Technology Acceptance Model (TAM) established perceived usefulness and perceived ease of use as central drivers of user acceptance, an influential lens for understanding teacher and student uptake of educational platforms [10]. For system-level evaluation, many education technology studies map user feedback to recognized software quality frameworks; ISO/IEC 25010 provides widely used quality characteristics (e.g., functional suitability, usability, security, maintainability) for specifying and evaluating software systems [11]. Together, these works support the need to test digital reading assessment platforms not only for effectiveness (learning gains) but also for user perception and quality, directly aligning with studies that compare outcomes against traditional methods and assess perceived performance using quality standards.

Uncertainties remain about the effects of using digital tools in reading beyond paper and pen. In [12], the authors emphasize the importance of examining how the use of digital devices can affect cognitive processes, as well as how different digital features may either support or hinder reading comprehension.

III. Methodology

A. Procedures

This study employed a quasi-experimental pretest–posttest control group design to examine the effectiveness of the PEER-based assessment platform in one of the secondary schools in the northern part of the Philippines, which is the Delfin Albano High School (DAHS). The system evaluated in this study was developed by a separate development team and served as the intervention platform.

Two comparable groups of secondary school students participated in the study: one group utilized the PEER platform using their own mobile devices, while the other group continued with the school's existing paper-and-pencil reading assessment practices. Intact classes were assigned to either the control or experimental group, as random assignment of individual students was not feasible in the school setting. Teachers in the control group conducted reading assessments using their usual methods, whereas teachers in the experimental group implemented the PEER platform for reading proficiency assessment after their training with the developers.

A pretest was administered to both groups to establish baseline reading proficiency levels. This included linear tasks where students read text-based materials and then answered related comprehension questions or quizzes. It also included non-linear tasks where students worked with visuals like graphs and charts and answered questions that required them to interpret and analyze these materials. After a one-month implementation period, a posttest was conducted using the same types of activities to determine changes in reading performance.

To complement the quantitative findings, a survey was conducted among students, class advisers, and teachers who utilized the PEER platform to evaluate its perceived software quality. Respondents for the software evaluation were selected through convenience sampling. The assessment of user perception was guided by the ISO/IEC 25010 software quality model, focusing on functional suitability, usability, security, and maintainability.

Prior to data collection, respondents were informed of the purpose of the study and assured that their participation was voluntary. They were also informed of their right to withdraw at any stage of the study and that all responses would be treated with confidentiality and used solely for research purposes.

For data analysis, paired and independent samples t-tests were used to compare pretest and posttest scores within and between the control and experimental groups to determine whether observed differences in reading proficiency gains were

statistically significant, while Cohen's *d* was also used to determine the effect size of Reading Proficiency Gains. Survey data were analyzed using the weighted mean to describe perceived software quality. A two-factor ANOVA without replication was conducted to examine further the differences across respondents' ratings and software quality criteria.

B. Platform Under Study

The digital reading proficiency assessment platform called PEER served as the intervention platform that was built by a team of developers following the Agile methodology. Its development unfolded as a deliberate and iterative process focused on improving how reading proficiency is assessed in the locale of the study. Their work began by clarifying the goals of the system and understanding how different users, such as teachers, administrators, and students, would engage with both the web and mobile applications. Through consultations and use case analysis, essential features were identified and prioritized to support efficient assessment, progress monitoring, and meaningful feedback.

With these goals in place, attention shifted to designing interfaces that were simple, intuitive, and responsive to users' needs. Early layouts and interaction flows were refined to ensure that the web and mobile applications of the platform offered a smooth and accessible experience. System development then brought these designs to life by integrating assessment logic, data storage, and synchronized access between the web and mobile applications. Continuous testing was carried out throughout this phase to ensure stability, accuracy, and ease of use. Fig. 1 shows the Admin Dashboard of the PEER web application [13]. This serves as the main control center where administrators manage key academic data such as sections, faculty profiles, student records, enrollment, and reading materials. It also provides an overview of system statistics, including the total number of enrolled students and faculty members.

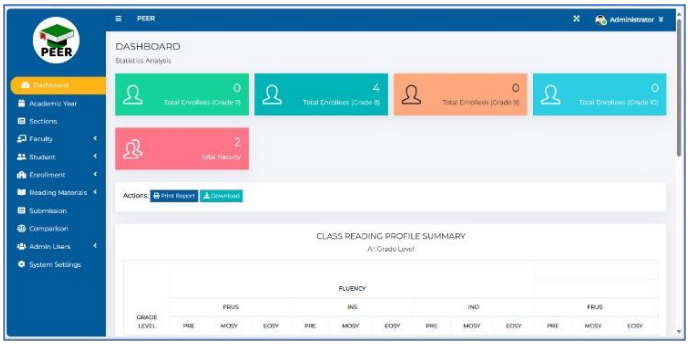


Fig. 1. Admin Dashboard of the PEER web App.

In addition, the dashboard features a class reading profile summary that allows users to quickly view and monitor student performance at a glance. Fig. 2 shows the Login Page of the PEER mobile application [13], which serves as the entry point for students before accessing the system's features. Through this interface, users are required to authenticate their accounts.

Fig. 3 shows the Reading Material Management, where teachers can upload and organize reading materials for student access, and Fig. 4 shows the Quiz Comprehension Management, which allows teachers to create and manage quizzes based on the selected reading material [13].

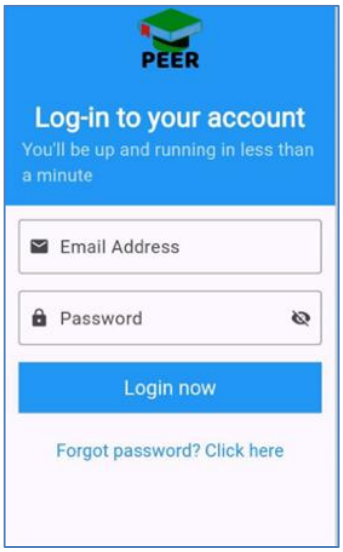


Fig. 2. Login Page of the PEER web app.

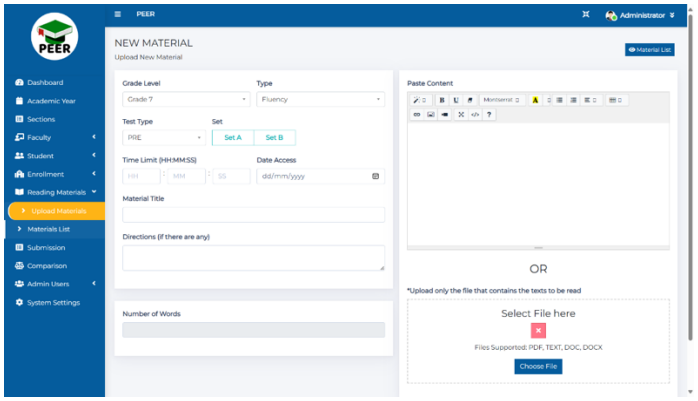


Fig. 3. Reading Material Management.

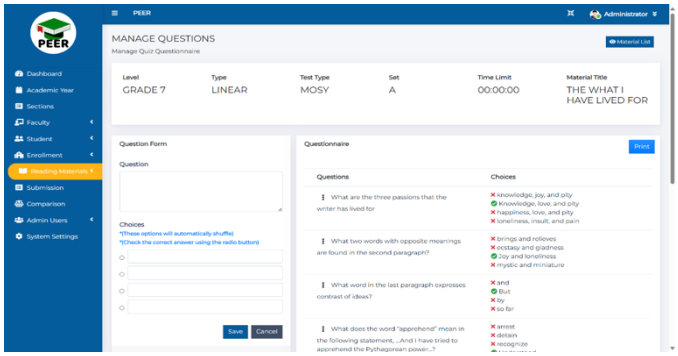


Fig. 4. Quiz Comprehension Management.

Fig. 5 shows the home page of the mobile application [13], including its simple and straightforward navigation menu that allows students to easily access the system's main features. Fig. 6, on the other hand, presents a sample recorded reading activity of a student, along with the quiz interface [13] where learners answer comprehension questions based on the assigned reading material.

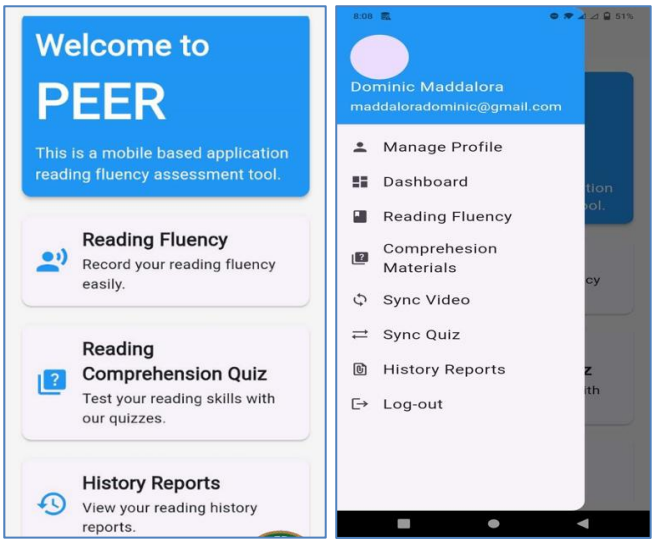


Fig. 5. Student portal.

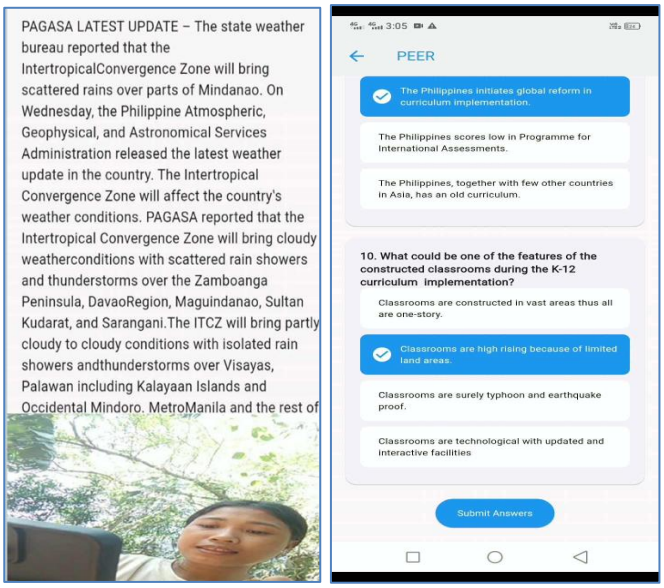


Fig. 6. Recorded student reading activity and quiz.

Following successful testing, the PEER platform was implemented by the developers in a public secondary school in Isabela, Philippines, and made available to first- and second-year high school students.

IV. Results

This section presents and discusses the findings of the platform's evaluation, focusing on a more constructive analysis of its effectiveness and user perception. The results are intended to provide empirical evidence to inform data-driven decisions regarding the potential adoption and wider implementation of PEER at the school level.

To examine the effectiveness of the PEER platform in comparison with traditional assessment methods, a t-test was employed as the primary statistical tool. Prior to conducting the t-test, the distribution of the data was assessed to ensure that the assumptions for parametric analysis were met. This assessment involved examining the skewness and kurtosis values of the scores.

As shown in Table I and Table II, the skewness values are close to zero, indicating that the score distributions are generally symmetric and did not exhibit substantial skewness. This suggests that the data were reasonably balanced around the mean. In addition, the kurtosis values are within acceptable ranges, further supporting the assumption that the data approximated a normal distribution. Taken together, these results indicate that the scores satisfied the normality assumption required for the application of t-tests.

TABLE I. Pre-Test and Post-Test Scores within Groups Across Linear and Non-Linear for the Control Group.

Measure of Data Distribution	Control Group			
	Pre-test Score		Post-test Score	
	Linear	Non-Linear	Linear	Non-Linear
Skewness	-0.041	-0.081	0.004	-0.022
Kurtosis	-0.174	-0.628	-0.899	-0.708

TABLE II. Pre-Test and Post-Test Scores within Groups Across Linear and Non-Linear for the Experimental Group.

Measure of Data Distribution	Experimental Group			
	Pre-test Score		Post-test Score	
	Linear	Non-Linear	Linear	Non-Linear
Skewness	-0.492	2.370	-0.674	-1.537
Kurtosis	-0.966	6.253	-0.447	0.961

A. Test of Difference Between the Pre- and Post-Test Scores Within Groups

After establishing that the data met the assumption of normality, a paired t-test comparing pre-test and post-test scores for both linear and non-linear reading assessments was computed.

For the control group, results presented in Table III indicate little to no improvement in reading performance. Mean scores in linear reading slightly declined from 4.067 in the pre-test to 3.767 in the post-test, with post-test scores showing slightly less variability (variance of 1.22 compared to 1.86 in the pre-test). The weak positive correlation between pre-test and post-test scores ($r = 0.217$), together with a non-significant p-value ($p = 0.300$), suggests that the observed changes were not statistically meaningful. These findings indicate that the traditional assessment approach did not lead to measurable improvements in students' reading performance. On the other hand, non-linear reading scores in the control group increased modestly from a mean of 4.400 to 4.933. Although the change reached statistical significance ($t = –2.570$, $p = 0.016$), the gain was small, indicating limited improvement using the traditional approach.

In contrast, the experimental group, as shown in Table IV, demonstrated substantial and statistically significant gains after using the PEER platform. Mean linear reading scores increased markedly from 4.567 to 8.033, while non-linear reading scores rose from 5.167 to 9.033. In addition to higher post-test means, the reduced variance in post-test scores reflects more consistent

performance among students. Although the correlations between pre-test and post-test scores were weak (linear: r = 0.085; non-linear: r = 0.181), the extremely small p-values ($p = 2.099 \times 10^{-9}$ for linear and $p = 9.004 \times 10^{-12}$ for non-linear) clearly indicate that these improvements were statistically significant.

TABLE III. RESULTS OF THE PAIRED T-TEST FOR PRE-TEST AND POST-TEST SCORES WITHIN CONTROL GROUPS.

	CONTROL GROUP			
	Linear		Non-Linear	
	Pre	Post	Pre	Post
Mean	4.067	3.767	4.400	4.933
Variance	1.857	1.220	1.628	1.237
Pearson Correlation	0.217		0.554	
t-stat	1.055		-2.570	
P(T<=t) one-tail	0.150		0.008	
t Critical one-tail	1.699		1.699	
P(T<=t) two-tail	0.300		0.016	
t Critical two-tail	2.045		2.045	

TABLE IV. RESULTS OF THE PAIRED T-TEST FOR PRE-TEST AND POST-TEST SCORES WITHIN EXPERIMENTAL GROUPS.

	EXPERIMENTAL GROUP			
	Linear		Non-Linear	
	Pre	Post	Pre	Post
Mean	4.567	8.033	5.167	9.033
Variance	3.357	2.033	1.868	2.723
Pearson Correlation	0.085		0.181	
t-stat	-8.537		-10.901	
P(T<=t) one-tail	1.049E-09		4.502E-12	
t Critical one-tail	1.699		1.699	
P(T<=t) two-tail	2.099E-09		9.004E-12	
t Critical two-tail	2.045		2.045	

Taken together, the results indicate that students assessed using the PEER platform achieved significantly greater gains in both linear and non-linear reading proficiency than those assessed using traditional methods, highlighting the platform's effectiveness as a digital assessment tool.

B. Test of Difference on the Reading Proficiency Gains Between the Control and Experimental Groups

Table V presents the mean gain scores in linear and non-linear reading assessments for both the control and experimental groups. The results show a clear contrast between the two groups in terms of reading proficiency gains. For the control group, the mean gain score in linear reading was -0.30, indicating a slight decline in performance from pre-test to post-test. In non-linear reading, the control group exhibited only a modest improvement, with a mean gain score of 0.53. These findings suggest that the traditional paper-and-pencil assessment approach resulted in little to no meaningful improvement in students' reading proficiency over the duration of the study.

In contrast, the experimental group demonstrated substantial gains in both reading assessment types. The mean gain score for linear reading was 3.47, while non-linear reading showed an even higher gain of 3.87. Independent samples t-test results revealed that the differences in gain scores between the control and experimental groups were statistically significant for both linear ($p = 2.88737 \times 10^{-10}$) and non-linear reading ($p = 3.99054 \times 10^{-11}$). These extremely small p-values provide strong evidence that the observed improvements in the experimental group were not due to chance but were associated with the use of the PEER platform.

TABLE V. MEAN GAIN SCORES FOR CONTROL AND EXPERIMENTAL GROUP.

	CONTROL GROUP		EXPERIMENTAL GROUP	
	Linear	Non-Linear	Linear	Non-Linear
MEAN GAIN SCORE	-0.30	0.53	3.47	3.87
p-value for Linear	2.88737E-10			
p-value for Non-Linear	3.99054E-11			

n=30

To determine the magnitude of the differences between groups, Cohen's *d* was computed using gain scores. Results revealed a very large effect size for both linear reading (d = 1.95) and non-linear reading (d = 2.02). These findings indicate that the use of the PEER platform produced a substantial and practically significant improvement in reading proficiency compared to traditional assessment methods (see Table VI).

TABLE VI. EFFECT SIZE (COHEN'S D) OF READING PROFICIENCY GAINS BETWEEN CONTROL AND EXPERIMENTAL GROUPS

Reading Assessment Type	Mean Gain		Cohen's d	Interpretation
	Control	Experimental		
Linear	-0.30	3.47	1.95	Very Large Effect
Non-Linear	0.53	3.87	2.02	Very Large Effect

The findings imply that the PEER-based assessment approach was significantly more effective than traditional methods in improving students' reading proficiency. The larger gains observed in the experimental group may be attributed to the platform's structured, consistent, and technology-supported assessment process, which allowed for timely monitoring of student performance and more focused engagement with reading tasks. Moreover, the greater gains in non-linear reading suggest that digital assessment environments may be particularly effective in supporting skills related to interpreting and integrating information from non-sequential texts. These results reinforce the value of integrating digital assessment platforms as part of reading proficiency programs to achieve stronger and more consistent learning outcomes.

C. Evaluation on the Performance of PEER

An evaluation was conducted using a survey questionnaire among the respondents to determine their perception with

respect to the functional suitability, usability, security and maintainability of the developed platform.

For the web application, all quality criteria received mean ratings above 4.50, corresponding to a descriptive rating of *Strongly Agree*. Functional suitability obtained a mean score of 4.72, suggesting that users perceived the system's features as well-aligned with their assessment needs. Usability also received a high mean of 4.69, indicating that users found the interface intuitive and easy to navigate. Similarly, security and maintainability were rated highly, with mean scores of 4.73 and 4.67 respectively, reflecting user confidence in data protection measures and the system's ability to support updates and modifications without difficulty.

Comparable results were observed for the mobile application, which also achieved *Strongly Agree* ratings across all criteria. Functional suitability (M = 4.57) and usability (M = 4.56) were particularly well rated, highlighting the effectiveness of the mobile app in supporting reading assessment tasks and providing a user-friendly experience. Security and maintainability each recorded mean scores of 4.52 and 4.55, respectively, indicating that users felt their data were secure and that the application was reliable and adaptable to future needs.

The overall mean scores of 4.70 and 4.55 for web and mobile platforms, respectively, underscore the consistency of user satisfaction across deployment environments. These findings suggest that the PEER platform meets high software quality standards and is perceived as reliable, secure, and easy to use by its users. The consistently strong ratings across all dimensions support the platform's suitability for broader implementation and reinforce its potential as an effective digital tool for reading proficiency assessment (see Table VII).

TABLE VII. PERCEIVED PERFORMANCE EVALUATION OF PEER

SW Quality Criteria	Web App		Mobile App	
	Mean	SD	Mean	SD
Functional Suitability	4.72	0.30	4.57	0.45
Usability	4.69	0.30	4.56	0.50
Security	4.73	0.24	4.52	0.47
Maintainability	4.67	0.31	4.55	0.48
Overall Mean	**4.70**		**4.55**	**0.49**

n=50

A two-factor ANOVA without replication was conducted to examine differences across respondents' ratings and software quality criteria. The computed values on Table VIII and Table IX revealed a statistically significant effect of respondents (for web: $F = 8.92, p < 0.001$ and for mobile: $F = 28.50, p < 0.001$), indicating that there were significant differences in the ratings provided by respondents. This suggests variability in how individuals evaluated the system. In contrast, the effect of software quality criteria (Functional Suitability, Usability, Security, and Maintainability) was not statistically significant, (for web: $F = 1.63, p = 0.202$ and for mobile: $F = 0.80, p = 0.45$). The findings indicate that while respondent evaluations on the web and mobile applications significantly varied, there were no statistically significant differences among the assessed software quality criteria. This suggests that the system demonstrates a consistent level of quality across all evaluated dimensions.

TABLE VIII. TWO-FACTOR ANOVA (WEB APP)

Source of Variation	SS	df	MS	F	P-value	F crit
Respondents	9.52	49	0.19	8.92	5.02E-20	1.48
SW Quality	0.07	2	0.04	1.63	0.201699	3.09
Error	2.14	98	0.02			
Total	11.73	149				

TABLE IX. TWO-FACTOR ANOVA (MOBILE APP)

Source of Variation	SS	df	MS	F	P-value	F crit
Respondents	32.25	49	0.656	28.50	1.49E-40	1.48
SW Quality	0.037	2	0.02	0.80	0.452383	3.09
Error	2.26	98	0.02			
Total	34.55	149				

V. DISCUSSION

This study sought to examine whether a digital reading assessment platform could improve students' reading proficiency compared with conventional paper-based assessment practices. The results provide encouraging evidence that the PEER platform contributed to measurable improvements in students' reading performance while also receiving strong acceptance from users. When viewed together, the statistical findings and user evaluations suggest that technology-supported assessment systems may play a meaningful role in strengthening reading evaluation practices in schools. However, it is worth noting that the study only involved a relatively small sample size, which may limit how far the findings can be applied to a wider population. Because of this, future studies are encouraged to test the platform with a larger and more diverse group of participants to further strengthen and confirm its effectiveness.

The results of the pre-test and post-test comparison revealed a clear contrast between the control and experimental groups. Students who participated in the traditional paper-and-pencil assessment showed little change in their performance across the two testing periods. In linear reading, the slight decrease in the mean score suggests that repeated exposure to similar reading tasks alone may not necessarily lead to improvement. Although the control group demonstrated a modest increase in non-linear reading scores, the magnitude of the improvement remained relatively small. This outcome reflects a limitation often discussed in assessment literature, where conventional testing approaches tend to emphasize measurement rather than actively supporting the learning process [14].

A different pattern emerged among students who used the PEER platform. Their post-test scores increased substantially in both linear and non-linear reading assessments, indicating that the digital platform may have created a more supportive environment for completing reading tasks. One possible explanation is that digital systems introduce elements of

interaction and immediacy that are typically absent in static assessment formats. When learners engage with reading materials through digital interfaces, the experience often becomes more dynamic and structured, which may encourage sustained attention and participation. Research on digital learning environments has consistently highlighted the role of interactive technologies in promoting learner engagement and facilitating deeper processing of instructional content [15].

The analysis of gain scores provides further insight into the observed improvements. Students in the experimental group achieved considerably larger gains compared with those in the control group. Because both groups were exposed to the same reading content and assessment format, the improvement cannot simply be attributed to familiarity with the test items. Rather, the results suggest that the structured digital environment of the PEER platform played an important role in supporting learning during the assessment process. Digital assessment tools often allow educators to organize tasks more effectively and track student progress in real time. Such features have been associated with improved formative assessment practices, as they enable teachers to monitor student performance more closely and make timely instructional adjustments [16]; [17].

Another notable finding relates to the improvement in non-linear reading proficiency. Students in the experimental group demonstrated slightly higher gains in this area compared with linear reading. This result is particularly relevant in light of contemporary literacy practices, where readers frequently encounter information presented in non-sequential formats such as charts, tables, infographics, and hyperlinks. Digital platforms naturally support these types of information structures, which may help learners develop the skills needed to interpret and synthesize complex sources. Recent studies on digital literacy emphasize that reading in online environments requires additional competencies beyond traditional comprehension, including the ability to locate, evaluate, and integrate information from multiple digital formats [18]; [19]. The PEER platform may therefore provide an assessment context that more closely mirrors the reading environments students encounter outside the classroom.

Beyond learning outcomes, the evaluation of the PEER platform revealed consistently high ratings across several software quality dimensions. Respondents strongly agreed that both the web and mobile applications demonstrated functional suitability, usability, security, and maintainability. These results are important because the success of educational technologies often depends not only on their instructional effectiveness but also on their usability and reliability in real classroom settings. Systems that are perceived as intuitive and easy to navigate are more likely to be adopted and sustained by teachers and students [20].

The positive evaluation of system security also deserves attention. In digital learning environments, the protection of student data is a critical concern for schools and administrators. The favorable ratings suggest that users trusted the platform's mechanisms for managing and storing information. At the same time, the strong perception of maintainability indicates that the system is flexible enough to accommodate updates and improvements over time. This adaptability is an important characteristic of educational software, as technological tools must evolve alongside changes in curriculum, pedagogy, and institutional needs [21]. Recent discussions in educational technology research further emphasize that scalable and adaptable digital systems are essential for supporting long-term integration in schools [22].

Taken together, the findings of this study highlight the potential value of digital assessment platforms in improving reading proficiency and enhancing the assessment experience. The significant gains observed among students who used the PEER platform, combined with the positive user perceptions of its quality and usability, suggest that such systems can serve as effective complements to traditional assessment methods. Rather than replacing conventional approaches entirely, digital platforms may offer additional opportunities for monitoring student progress, providing more engaging assessment experiences, and supporting data-informed decision-making in the classroom.

As schools continue to integrate technology into teaching and learning, the role of digital assessment tools is likely to become increasingly important. Platforms such as PEER demonstrate how technology can be used not only to measure learning outcomes but also to create assessment environments that align more closely with the digital contexts in which students read and interact with information. With thoughtful implementation, systems like PEER may contribute to more responsive and effective reading proficiency programs in secondary education.

VI. Conclusion

This study highlights how thoughtfully designed digital platforms can meaningfully improve the way reading proficiency is assessed in school settings. By streamlining assessment procedures and organizing results in a systematic manner, digital tools help schools establish more reliable and efficient reading comprehension assessment practices. More importantly, they allow teachers to spend less time on repetitive, manual tasks and more time interpreting results and responding to learners' needs. In this way, digital assessment platforms do not simply modernize existing practices; they strengthen the instructional support system that underpins reading development.

At a broader level, the findings suggest that technology-supported assessment is well aligned with the changing nature of reading literacy. As students interact with a wider range of texts, spanning linear, non-linear, and multimodal formats, digital platforms provide a flexible yet standardized approach to evaluating reading skills. The favorable evaluation of software quality across platforms also implies that when digital assessment tools are built in accordance with established quality standards, they are more likely to be accepted by users and sustained over time within educational institutions.

Looking ahead, future studies may build on this work by investigating how continued use of digital reading assessment platforms influences literacy growth over longer periods and across different grade levels. Expanding the scope of participants and implementing the system in multiple school contexts could further strengthen the applicability of the results.

There is also room to explore more advanced features, such as adaptive assessments, learning analytics, or AI-driven feedback mechanisms, to support more personalized reading interventions. Finally, qualitative inquiries into teachers' classroom practices and students' learning experiences would offer valuable insights into how assessment data from digital platforms are translated into concrete instructional improvements and meaningful learning outcomes.

ACKNOWLEDGMENT

The author sincerely thanks the teachers and students of Delfin Albano High School, under the leadership of Dr. Orlyne T. Demerin, for their support and valuable inputs. Appreciation is extended to the project team, namely Revilyn Totto, Christine Rondero, Vanessa Pagauitan, and Julius Lemu, for sharing PEER and their study on the developed platform.

REFERENCES

[1] N. B. Acedillo, "Improving the reading comprehension skills of Grade 5 pupils through contextualized learning materials: A school-based research," International Journal of Scientific and Research Publications, vol. 13, no. 3, pp. 380–392, Mar. 2023, doi: 10.29322/IJSRP.13.03.2023.p13545.

[2] OECD, "Programme for International Student Assessment (PISA)," Organisation for Economic Co-operation and Development. [Online]. Available: https://www.oecd.org/en/about/programmes/pisa.html

[3] J. Hu and Y. Wang, "Influence of students' perceptions of instruction quality on their digital reading performance in 29 OECD countries: A multilevel analysis," Computers & Education, vol. 189, p. 104591, 2022, doi: 10.1016/j.compedu.2022.104591.

[4] C.-M. Chen et al., "A web-based collaborative reading annotation system with gamification mechanisms to improve reading performance," Computers & Education, vol. 144, p. 103697, 2020.

[5] P. Auphan, A. Magnan, and J. Ecalle, "The high potential of computer-based reading assessment," Canadian Journal of Learning and Technology, vol. 46, no. 3, 2020.

[6] Chen, Danlei, and Gale Macleod. 2021. "Effectiveness of Digital Tools to Support Pupils' Reading in Secondary School: A Systematised Review." International Journal of Mobile and Blended Learning 13, no. 2 (January): 1–16. https://doi.org/10.4018/IJMBL.2021040101

[7] P. Delgado, C. Vargas, R. Ackerman, and L. Salmerón, "Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension," Educational Research Review, vol. 25, pp. 23–38, 2018.

[8] P. M. Stecker, L. S. Fuchs, and D. Fuchs, "Using curriculum-based measurement to improve student achievement: Review of research," Psychology in the Schools, vol. 42, no. 8, pp. 795–819, 2005, doi: 10.1002/pits.20113.

[9] N. Förster and E. Souvignier, "Learning progress assessment and goal setting: Effects on reading achievement, reading motivation, and reading self-concept," Learning and Instruction, vol. 32, pp. 91–100, 2014, doi: 10.1016/j.learninstruc.2014.02.002.

[10] 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, doi: 10.2307/249008.

[11] ISO/IEC, ISO/IEC 25010:2011 Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—System and software quality models. Geneva, Switzerland: International Organization for Standardization, 2011.

[12] Altamura, Lidia. 2024. "Turning the Page: What Research Indicates About Print vs. Digital Reading." Research Review 2, no. 4 (November 4, 2024 CDT).

[13] Totto, R., Rondero, R., Pagauitan, V., Lemu, J., and Tarun, I. 2024. PEER: A Web and Mobile Platform for Efficient Evaluation of Reading Proficiency. Unpublished capstone project.

[14] P. Black and D. Wiliam, "Assessment and classroom learning," Assessment in Education: Principles, Policy & Practice, vol. 5, no. 1, pp. 7–74, 1998.

[15] R. C. Clark and R. E. Mayer, E-learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning, 4th ed. Hoboken, NJ, USA: Wiley, 2016.

[16] D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice," Studies in Higher Education, vol. 31, no. 2, pp. 199–218, 2006.

[17] C. Redecker and Ø. Johannessen, "Changing assessment—Towards a new assessment paradigm using ICT," European Journal of Education, vol. 56, no. 1, pp. 79–96, 2021.

[18] D. J. Leu, E. Forzani, C. Rhoads, C. Maykel, C. Kennedy, and N. Timbrell, "The new literacies of online research and comprehension," Educational Psychologist, vol. 50, no. 1, pp. 37–59, 2015.

[19] P. Delgado, C. Vargas, R. Ackerman, and L. Salmerón, "Don't throw away your printed books: A meta-analysis on the effects of reading media on comprehension," Educational Research Review, vol. 35, p. 100433, 2022.

[20] J. Nielsen, Usability Engineering. San Francisco, CA, USA: Morgan Kaufmann, 2012.

[21] R. S. Pressman and B. R. Maxim, Software Engineering: A Practitioner's Approach, 9th ed. New York, NY, USA: McGraw-Hill, 2020.

[22] M. Bond, K. Buntins, S. Bedenlier, O. Zawacki-Richter, and M. Kerres, "Mapping research in student engagement and educational technology in higher education: A systematic evidence map," International Journal of Educational Technology in Higher Education, vol. 20, no. 1, pp. 1–21, 2023.

Hybrid Learning-to-Rank Approach for Complex Information Retrieval Systems

Application to Biomedical Question Answering

Fatma Zohra Bessai-Mechmache[1], Yasmine Hanifi[2], Damia Lyna Ait Idir[3]
CERIST Research Center, Algiers, Algeria[1]
Computer Science Department, University of Algiers, Algiers, Algeria[2, 3]

***Abstract*—Biomedical question answering presents significant challenges due to the complexity of biomedical language and the need for precise information retrieval. This study aims to improve the performance of a biomedical information retrieval system through a hybrid learning-to-rank framework. Specifically, we combine lexical (BM25) and semantic (BioBERT) representations to form hybrid inputs for RankFormer, a transformer-based ranking model. This hybrid representation captures both surface-level term matching and deep contextual understanding. Experiments conducted on the BioASQ dataset show that our approach achieves better ranking performance compared to the standalone lexical or neural baselines, reaching a MAP@10 of 0.9614 and an nDCG@10 of 0.9320. These results highlight the effectiveness of hybrid input representations in enhancing biomedical answer ranking.**

***Keywords*—Learning-to-rank; information retrieval; hybrid learning-to-rank; transformer-based ranking model; biomedical question answering**

I. Introduction

Ranking in large-scale information retrieval can be seen as a multi-stage process where efficiency, relevance, and ranking stability must be jointly addressed.

Biomedical question answering (BQA) sits at the intersection of natural language processing and information retrieval (IR), aiming to return relevant answers to complex biomedical questions. The volume of biomedical literature continues to grow exponentially, and the terminology itself is often highly specialized and ambiguous. Together, these factors tend to make effective information retrieval genuinely challenging in real-world biomedical settings.

Traditional lexical retrieval models, with BM25 as the most common example, rely on exact term matching and have, for years, served as solid baselines in information retrieval. They are fast, relatively simple to deploy, and generally do a good job when surface-level word overlap is enough. However, such models tend to struggle with semantic variation, including synonymy and context-dependent meanings, all of which appear frequently in biomedical writing.

Neural language models like BioBERT take a different direction by using contextualized representations that aim to capture semantic relationships beyond lexical similarity. Although these models have demonstrated strong performance in biomedical natural language understanding tasks, they can underperform when precise term matching matters. They typically require computationally intensive re-ranking.

Recent studies suggest that mixing lexical and semantic signals can alleviate the limitations of individual models. Hybrid retrieval strategies that use sparse (exact relevance signal) with dense (contextual relevance signal) representations often perform better across multiple benchmarks, and the gains tend to be more noticeable in specialized domains. In the biomedical domain, where both precise terminology and deep contextual understanding are crucial, such strategies are particularly relevant.

In this work, we aim to improve the performance of biomedical information retrieval systems by adopting a hybrid learning-to-rank approach. Specifically, we propose a method that brings together lexical scores from BM25 and contextual embeddings derived from BioBERT to form a hybrid input for RankFormer, a transformer-based listwise ranking model. Rather than treating sparse and dense relevance signals separately, the model is trained to learn from them jointly within a supervised setting, which appears to encourage rankings that are not only more accurate but also more sensitive to meaning and context. We experiment with the proposed approach on the BioASQ Task B 2024 dataset, where the results demonstrate that combining hybrid representations with listwise learning can lead to noticeable gains over both purely lexical methods and standalone neural baselines.

II. Related Work

Information retrieval is the process of locating information items that best align with a user's query [1], [2]. In biomedical contexts, this task becomes especially demanding. The huge volume of biomedical literature, along with its domain-specific terminology, means that retrieval systems often play a practical role in clinical decision support and day-to-day biomedical research [3]. Not surprisingly, a wide range of methods has been proposed over time to improve retrieval quality, from classic lexical techniques to more recent neural models and, increasingly, hybrid approaches that try to balance precision with semantic understanding.

A. Lexical Methods (Sparse Retrieval Methods)

Traditionally, lexical approaches relying on word overlap have been the foundation of information retrieval [4]. Sparse retrieval methods use inverted indexes to link terms to documents, which helps explain why they remain

computationally efficient even at a large scale. Over the years, researchers have proposed a range of weighting and normalization schemes, rising to multiple Term Frequency-Inverse Document Frequency (TF-IDF) models [5]. Because it offers a reasonable balance between effectiveness and simplicity, BM25 [6] is one of the sparse models that is often viewed as the default choice in real-world systems. In practice, it processes documents as bags of words and assigns scores based on the presence of query terms independently of their position or proximity in the text [7].

A key strength of BM25 lies in its ability to normalize document length and address query term saturation. This enables it to avoid bias toward either very short to very long documents and to reduce the disparate influence of excessively repeated terms. BM25 has been extended to incorporate multiple document fields and specify how their individual scores should be combined [8].

Lexical models continue to be popular largely because they are fast and easy to deploy at scale. In many general-purpose retrieval settings, lexical methods are usually enough. However, when the language becomes more specialized, as in biomedical texts, synonymy and contextual ambiguity are common, and methods based solely on word overlap are likely to miss relevant documents even when the underlying concepts align [9].

B. Semantic Methods (Dense Retrieval Methods)

Efforts to improve information retrieval systems have increasingly turned toward neural-based semantic retrieval approaches. Early work in this category relied on convolutional and recurrent neural networks to represent textual data. In practice, both queries and documents are mapped into a shared vector space assumed to capture more nuanced representations of meaning. Relevance is then estimated through similarity functions trained to reflect semantic alignment between these representations [10], [11].

Additionally, the neural models benefit from their ability to learn non-linear text representations, which explain why they often outperform traditional lexical approaches [11]. A more noticeable advancement in the field came with the introduction of transformer models. Transformer-based models such as BERT [12] and GPT [13], thanks to the attention mechanism, have significantly improved retrieval performance by enabling the model to focus attention on the most relevant parts of both queries and documents during the matching process [14].

Neural models such as BioBERT were introduced to move beyond surface-level term matching and toward a richer notion of semantic similarity. As a domain-adapted version of BERT, BioBERT is pre-trained on large-scale biomedical texts. This model has reported strong performance on biomedical question answering and document ranking benchmarks [15]. Neural models offer an improved ability to relate concepts that are not lexically identical. For BQA systems, where relevant evidence is often paraphrased rather than repeated verbatim, this contextual sensitivity is crucial [16].

C. Hybrid Methods

The development of hybrid approaches is motivated by the strengths and complementary results produced by sparse and dense retrieval models [1]. Hybrid methods typically merge lexical and semantic scores using fusion techniques. Among the most widely used strategies are Convex Combination and Reciprocal Rank Fusion (RRF).

Although both strategies are widely used, Convex Combination appears to be the more flexible option in many settings. It requires only minimal score normalization and, according to prior studies, has in several cases outperformed RRF [17]. One practical advantage, however, is that it tends to be sample-efficient, adapting reasonably well to specific domains with relatively few examples.

A common strategy involves using BM25 to retrieve an initial set of candidate documents, which are then re-ranked using a dense encoder such as BioBERT. This two-stage architecture has been shown to significantly improve retrieval performance in various QA benchmarks. More recent work has pushed this integration further by directly fusing lexical and dense scores.

D. Learning-to-Rank

Learning-to-rank (LTR) is designed as a supervised learning approach for ordering documents by how relevant they are to a given query. LTR has become an important component of modern information retrieval systems, especially in settings where simple heuristics no longer suffice. That said, LTR actually covers a range of modeling choices rather than a single, unified method.

Most LTR techniques are commonly organized into three broad paradigms [18], each of which frames the ranking problem in a slightly different way. The pointwise approach, for instance, treats the ranking problem as a classification task, where each document is assigned a relevance score independently of the others. This setup is attractive because it is easy to implement and fits neatly within existing machine learning pipelines. However, by scoring documents independently, it arguably misses a key aspect of ranking because relevance is often relative and emerges from direct competition among documents shown for the same query.

Pairwise methods try to close this gap by shifting the learning objective toward preferences between pairs of documents. Instead of asking whether a document is relevant in absolute terms, the model learns which of two documents should be ranked higher. RankNet, introduced by Burges et al. [18], is a well-known example; it uses a neural network to estimate the probability that one document should be ranked ahead of another.

Listwise methods take a more holistic view by considering the entire set of documents associated with a query during training. One of the more influential examples is ListNet [19]. ListNet defines a probability distribution over possible permutations of the result list. This formulation appears to let the model capture interactions among documents more directly.

Recent progress in deep learning has pushed learning-to-rank models toward transformer-based approaches, largely due to their ability to capture complex contextual relationships. One such transformer-based listwise LTR model that incorporates traditional listwise objectives with innovative listwide assessment objectives is RankFormer [20]. This design appears

to help the model capture dependencies across an entire document list, leading to improved ranking performance.

Our work follows this line of research but takes a different angle. We propose a hybrid approach in which both BM25 scores and BioBERT embeddings are provided as input signals to a RankFormer-based re-ranker. Rather than treating lexical retrieval and semantic re-ranking as two clearly separated stages, we integrate sparse and dense features from the outset. The idea here is that BioBERT captures domain-specific semantics, whereas BM25 tends to anchor the ranking in exact term matches. In order to take benefit of the complementary advantages of both sparse and dense retrieval for more precise biomedical passage ranking, the two are combined early in the ranking process.

III. Methodology

The proposed methodology adopts a hierarchical learning-to-rank framework to progressively enhance document relevance for biomedical question answering. To take advantage of complementary relevance signals at different stages of the ranking process, the approach combines supervised listwise learning with semantic re-ranking and lexical retrieval. Three sequential steps are used to demonstrate the workflow: 1) BM25 is used for initial candidate retrieval; 2) BioBERT is used for semantic re-ranking; and 3) a hybrid RankFormer model is used for final ranking.

This multi-stage strategy seeks to balance efficiency and effectiveness by leveraging the strong recall of sparse retrieval with the semantic sensitivity of dense representations and the optimization capabilities of learning-to-rank models.

For a given query, we start by keeping the top 50 documents returned by BM25. At this stage, the goal is coverage rather than finesse. BioBERT then re-ranks this shortlist, filtering it down to the 25 documents that appear to align most closely with the query at a semantic level. These candidates are then passed to RankFormer, which generates the final top-10 ranking. This multi-stage ranking strategy has proven to be effective in advanced information retrieval systems, as it brings together the efficiency of sparse retrieval with the precision of dense methods and learning-to-rank techniques [21].

A. Dataset

We carry out our experiments on the BioASQ Task B 2024 dataset, a reference benchmark for biomedical question answering. The dataset consists of expert-formulated biomedical questions. Each question is linked to a set of supporting materials: relevant PubMed documents, text snippets, biomedical concepts, and expert answers curated by domain specialists [22].

BioASQ Task B queries are natural-language biomedical questions, categorized into four classes: Factoid questions (single entity answers), List questions (multiple entity answers), Yes/No questions, and Summary (ideal answer) questions.

For retrieval and ranking purposes, the dataset is preprocessed into three components:

- Queries: Question texts are extracted from the original JSON files and stored in JSONL format, where each entry contains a unique query identifier and the query body.
- Documents: Document collections are constructed by concatenating snippets associated with the same source URL, resulting in a dense textual corpus indexed using the Pyserini [23] toolkit.
- Relevance Judgments (Qrels): Ground-truth relevance annotations are formatted according to the standard TREC qrels specification, linking each query to relevant documents.

This preprocessing task ensures compatibility with both sparse and dense retrieval models and provides a reliable evaluation framework for learning-to-rank experiments.

B. Initial Lexical Retrieval with BM25

The suggested methodology starts with BM25, a probabilistic sparse retrieval model that is frequently used in information retrieval because of its interpretability and robustness. BM25 ranks documents based on exact query term matching using term frequency saturation and document length normalization.

We test four distinct BM25 configurations in order to examine the effects of BM25 hyperparameters. Precisely, we vary the term frequency parameter "k_1" and the length normalization parameter "b":

- $k_1 = 0.9$, $b = 0.4$
- $k_1 = 0.9$, $b = 0.8$
- $k_1 = 1.2$, $b = 0.4$
- $k_1 = 1.2$, $b = 0.8$

For each query, the top 50 documents are retrieved from the indexed corpus. This candidate set is intentionally kept broad in order to maximize recall, ensuring that potentially relevant documents are preserved for subsequent re-ranking stages.

C. Semantic Re-ranking with BioBERT

In the second stage, we move beyond lexical matching and introduce semantic re-ranking with BioBERT, a transformer model pre-trained on large-scale biomedical datasets. The idea is that once BM25 has gathered a set of plausible candidates, BioBERT reconsiders them through a contextual goal. Instead of relying purely on term overlap, it can account for synonymy and the kinds of domain-specific expressions that often appear in biomedical writing.

At first, we applied BioBERT in its pre-trained form, using it to score the top 50 documents retrieved by BM25 for each query. Nevertheless, initial tests show that task sensitivity is low in this zero-shot setting. The model is fine-tuned on the BioASQ relevance judgments in order to overcome this constraint.

Fine-tuning data are constructed by pairing queries with candidate documents and assigning binary relevance labels derived from the qrels. A subset of 1,000 query-document pairs is randomly sampled and split into 80% for training and 20% for validation. Inputs are tokenized using the BioBERT tokenizer,

with truncation and padding applied to a maximum sequence length of 512 tokens.

The model is fine-tuned as a binary classifier for three epochs using a learning rate of 2e-5, a batch size of 8, and a weight decay of 0.01. Validation accuracy is used to monitor training progress. After fine-tuning, BioBERT demonstrates improved relevance prediction and is used to re-rank the BM25 candidate set. The top 25 documents per query are retained for the final ranking stage. This semantic re-ranking stage serves as a critical bridge between traditional lexical matching and transformer-based contextual understanding, yielding more precise retrieval results in the biomedical domain.

D. RankFormer Hybrid Learning-to-Rank Model

The final ranking stage employs RankFormer, a transformer-based listwise learning-to-rank model capable of modeling inter-document dependencies within a candidate list. Unlike conventional pipelines that treat lexical and semantic re-ranking as isolated steps, our approach integrates hybrid relevance signals directly into the RankFormer training process.

Each training instance is defined as a triplet (q, D, L), where q denotes a query, D = {d_1, d_2, ..., d_n} is the set of candidate documents, and L represents a list-level supervision signal referred to as the listwise label. Each document d_i is associated with a document ID, a textual representation, and a relevance label {0, 1}, a predicted score, and a rank reflecting its position within the list.

To generate input scores for RankFormer, we compute a normalized linear combination of BM25 scores and fine-tuned BioBERT scores. The hybrid score for a document is given by Eq. (1) [24]:

$$score = \alpha \cdot scorebm25 + (1 - \alpha) \cdot scorebiobert \quad (1)$$

where, $\alpha \in [0, 1]$ controls the relative contribution of lexical and semantic signals. In our experiments, α is empirically set to 0.5. The choice of setting α=0.5 is driven by the objective of achieving a balanced integration of heterogeneous scoring signals. In biomedical document retrieval, BM25 ensures precise lexical matching for highly specialized terminology, whereas BioBERT captures richer semantic context and synonymy. Assigning equal weights prevents the RankFormer input from being biased toward either surface-level keyword matching or purely latent representations, thereby maintaining the precision of the former while benefiting from the recall of the latter.

In addition to document-level relevance labels, RankFormer incorporates list-level supervision. A binary listwide label "L" is assigned to each query: the label is set to 1 if the candidate list contains at least three relevant documents, and 0 otherwise. This supervision enables the model to assess global ranking quality beyond individual document relevance.

To optimize the model, the training objective combines a listwise ranking loss and a listwide loss.

The listwise loss encourages correct relative ordering of documents within each query candidate list. It follows a Softmax-based ranking loss [19], inspired by ListNet and defined by Eq. (2) as follows:

$$Llistwise = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{|D_i|} (\log \sum_{k=j}^{|D_i|} \exp(s_{i,k}) - s_{i,j}) \quad (2)$$

where, S_{ij} denotes the predicted score for the j-th document in the i-th list and N is the total number of training queries.

The listwide loss is defined by Eq. (3) as a mean squared error (MSE) [25] between the predicted list quality and the true listwide label.

$$Llistwide = \frac{1}{N} \sum_{i=1}^{N} (l_i - \widehat{l_i})^2 \quad (3)$$

where, l_i is the true listwide label for the i-th query, and $\widehat{l_i}$ is the predicted list quality output by the model.

As specified in Eq. (4), the final training objective function is expressed as a weighted combination of both losses.

$$Ltotal = \alpha \cdot Llistwise + (1 - \alpha) \cdot Llistwide \quad (4)$$

where, α is set to 0.7, to prioritize accurate document ranking while maintaining sensitivity to list-level performance.

This hybrid supervision strategy allows RankFormer to jointly optimize fine-grained relevance ordering and holistic ranking performance, making it particularly suitable for biomedical information retrieval tasks.

The RankFormer-Hybrid model architecture is built upon a BioBERT backbone, which serves as a pre-trained encoder to extract domain-specific contextual embeddings. To manage overhead, these high-dimensional representations are projected into a 128-dimensional space via a linear transformation layer. The primary ranking operations are then conducted by a streamlined Transformer encoder consisting of two layers and four attention heads. Within this framework, query-document fusion is achieved by appending the query representation to each document vector, allowing the model to capture fine-grained, query-dependent relevance signals. The architecture concludes with a dual-level scoring mechanism, an individual linear layer computes local scores for each document, while a separate listwide score is generated by averaging document representations through a global linear layer.

As illustrated in Table I, the entire architecture is optimized using a weighted hybrid loss function, Eq. (4), which balances the precision of relative document ordering with the overall coherence of the retrieved list.

TABLE I. RANKFORMER-HYBRID MODEL ARCHITECTURE

Component	Description
Pre-trained encoder	BioBERT (dmis-lab/biobert-base-cased-v1.1)
Linear projection layer	Linear to a 128-dimensional space
Transformer encoder	2 layers, 4 attention heads, batch-first = true
Query-document fusion	Adding the query representation to each document
Score estimation	Individual linear layer for each document (score-layer)
Listwide score	Average of documents → linear layer (listwide-layer)
Learning loss	0.7· Loss-listwise + 0.3· Loss-listwide

E. Evaluation Metrics

Ranking performance is evaluated using standard information retrieval metrics at a cutoff of k = 10, consistent with the BioASQ Task B setting. The metrics include MAP@10, MRR@10, nDCG@10, Precision@10, and Recall@10. Together, these metrics provide a comprehensive assessment of ranking accuracy, early precision, and retrieval completeness.

A brief definition of these evaluation metrics [26] is as follows:

- Mean Average Precision at 10 (MAP@10) computes the average precision across all relevant documents within the top 10 retrieved, weighted by their position. This metric emphasizes both precision and the ranking order of relevant documents.
- Mean Reciprocal Rank at 10 (MRR@10) captures the position of the first relevant document in the ranking. It is defined as the reciprocal of the rank at which the first relevant document appears, providing a measure of how quickly a system can retrieve a relevant result.
- Normalized Discounted Cumulative Gain at 10 (nDCG@10) evaluates the quality of the ranking by assigning higher importance to relevant documents appearing higher in the list. It accounts for graded relevance and uses a logarithm discount factor to penalize lower-ranked relevant documents.
- Precision at 10 (P@10) measures the proportion of relevant documents among the top 10 retrieved items, offering an intuitive assessment of immediate result relevance.

Recall at 10 (R@10) quantifies the fraction of all relevant documents that are retrieved within the top 10 results, thereby reflecting the model's overall retrieval coverage.

These complementary metrics provide a comprehensive evaluation of the studied models, capturing different aspects of retrieval performance, including completeness, early precision, and ranking accuracy. Following the proposed methodology, the next sections present and analyze the experimental results obtained within this evaluation framework.

IV. Experimental Results and Discussion

The experimental results demonstrate a clear and consistent improvement in retrieval performance as the ranking pipeline evolves from lexical to dense and finally to supervised learning-to-rank approaches. This progression highlights the complementary nature of sparse and dense relevance signals and confirms the central role of supervised ranking in biomedical information retrieval.

A. BM25 Results

The BM25 baseline provides a strong lexical foundation, achieving competitive precision-oriented metrics when properly tuned. These results are consistent with prior findings showing that sparse retrieval models remain effective in biomedical settings when domain-specific terminology is well represented. However, the observed limitations in recall and nDCG indicate that lexical matching alone is insufficient to fully capture the semantic variability inherent in biomedical queries and documents.

Table II presents the evaluation results for the four configurations of the parameters k_1 and b across the five standard retrieval metrics at rank 10.

TABLE II. Evaluation Results of BM25 on the BioASQ Task B 2024 Dataset.

Metrics @10	k_1=0.9, b=0.4	k_1=0.9, b=0.8	k_1=1.2, b=0.4	k_1=1.2, b=0.8
MAP	0.6100	0.6255	0.6073	0.6263
MRR	0.8965	0.8866	0.8978	0.8886
Precision	0.5885	0.6044	0.5871	0.6088
Recall	0.4703	0.4860	0.4696	0.4884
nDCG	0.7761	0.7847	0.7744	0.7877

The results reveal that the configuration $k_1 = 1.2$, $b = 0.8$ consistently yields the best performance for most metrics. Interestingly, the highest MRR@10 is reached with $k_1 = 1.2$, $b = 0.4$, suggesting that while this setting helps retrieve the first relevant document at a higher rank, the overall ordering may be less optimal than with $b = 0.8$.

These findings emphasize the importance of hyperparameter tuning in lexical retrieval and confirm that BM25, even as a sparse model, can provide a strong and competitive baseline when properly optimized.

B. BioBERT Results

We evaluate the performance of BioBERT in two experimental settings: in its original pre-trained form and after being fine-tuned on the BioASQ TASK B 2024 training set. The results shown in Table III indicate a substantial improvement across all evaluation metrics after fine-tuning.

TABLE III. Evaluation Results of BioBERT Before and After Fine-Tuning.

Metrics	Pre-trained model	Fine-tuned model
MAP@10	0.6430	0.8180
MRR@10	0.7512	0.8883
Precision@10	0.5097	0.6088
Recall@10	0.3589	0.4884
nDCG@10	0.7578	0.8917

To further illustrate these differences, Fig. 1 presents a bar chart comparing the performance of the two models. The fine-tuned BioBERT clearly outperforms the pre-trained version, achieving a MAP@10 of 0.8180 compared to 0.6430, and an nDCG@10 of 0.8917 versus 0.7578. These results emphasize the effectiveness of domain-specific fine-tuning in enhancing ranking quality for biomedical information retrieval tasks.

To better assess the impact of dense retrieval methods in biomedical information retrieval, we compare the top-performing BM25 configuration with the fine-tuned BioBERT model. The results, summarized in Table IV, highlight the advantages of using a semantic approach over a traditional sparse retrieval model.

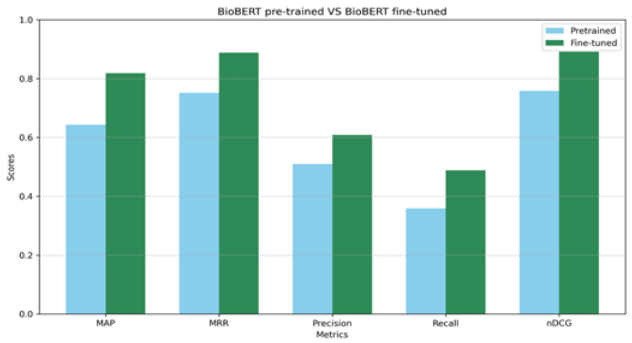


Fig. 1. Comparison of BioBERT performance in pre-trained and fine-tuned settings for each evaluation metric at rank 10.

TABLE IV. RETRIEVAL PERFORMANCE COMPARISON BETWEEN BM25 AND FINE-TUNED BIOBERT ON THE BIOASQ TASK B 2024 DATASET.

Metrics	BM25 (k_1 = 1.2, b = 0.8)	Fine-tuned BioBERT
MAP@10	0.6263	0.8180
MRR@10	0.8886	0.8883
Precision@10	0.6088	0.6088
Recall@10	0.4884	0.4884
nDCG@10	0.7877	0.8917

As illustrated in Fig. 2, BioBERT outperforms BM25 across most metrics, notably achieving improvements of +19.2% in MAP@10 and +10.4% in nDCG@10. These gains confirm the added value of contextualized language representations in capturing nuanced biomedical terminology and semantics, which are often missed by term-matching approaches like BM25.

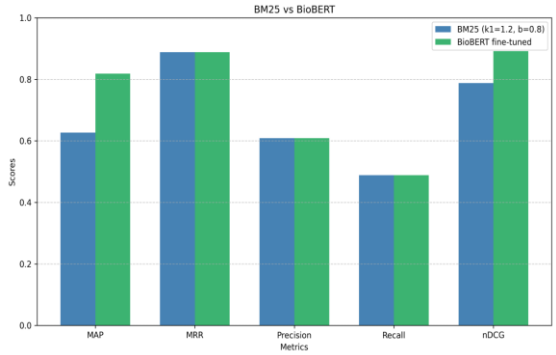


Fig. 2. Comparative performance of BM25 and fine-tuned BioBERT.

The introduction of BioBERT significantly improves ranking quality, particularly in terms of MAP@10 and nDCG@10. These gains can be attributed to the model's ability to encode contextual semantics, enabling it to capture synonymy, abbreviations, and implicit relationships that are not accessible to term-based approaches. Interestingly, while BioBERT and BM25 achieve comparable Precision@10 and MRR@10 scores, BioBERT consistently produces a more coherent ordering of relevant documents. This suggests that dense representations primarily enhance ranking quality rather than retrieval quantity, a pattern frequently observed in domain specific retrieval tasks.

C. RankFormer-Hybrid Model Results

Table V presents the performance of the proposed RankFormer-Hybrid model, summarizing results after 5 and 6 training epochs. These outcomes reflect the integration of hybrid relevance signals into the model's learning process.

TABLE V. RANKFORMER-HYBRID PERFORMANCE AT RANK 10 AFTER 5 AND 6 TRAINING EPOCHS.

Metrics / Epochs	5	6
MAP@10	0.9614	0.9526
MRR@10	0.9196	0.9118
Precision@10	0.6277	0.6242
Recall@10	0.8124	0.8081
nDCG@10	0.9320	0.9320

The model demonstrates outstanding performance across all evaluation metrics, even with a limited number of training epochs. At epoch 5, the model reaches a MAP@10 of 0.9614, MRR@10 of 0.9196, and nDCG@10 of 0.9320. These values remain stable or slightly decrease at epoch 6 due to the application of early stopping, which was employed to prevent overfitting and ensure optimal generalization.

Such high scores with minimal training suggest that the hybrid supervision significantly accelerates the convergence of the model by providing rich and informative ranking signals. This strong external guidance reduces reliance on long training schedules and enhances learning efficiency.

The rapid convergence observed within a small number of training epochs indicates that the hybrid supervision provides a stable and informative optimization signal. This behavior is particularly relevant for applied systems, where computational efficiency and training stability are critical design constraints.

These findings confirm the effectiveness of the proposed RankFormer-Hybrid model and highlight the value of incorporating complementary lexical and semantic signals in ranking models. This hybrid strategy leads to more nuanced and accurate rankings, particularly beneficial in complex systems such as biomedical information retrieval.

D. Comparative Analysis

The final comparison brings together the best-performing configurations of all evaluated models in order to highlight the progressive improvement achieved through increasingly sophisticated retrieval strategies. The results presented in Fig. 3 show a clear and consistent trend. The performance improves markedly as we move from traditional sparse retrieval to dense, and finally to learning-to-rank models.

BM25, a strong lexical baseline, performs well in precision-oriented metrics such as MAP and MRR, especially when its hyperparameters are finely tuned. However, it remains limited in its ability to capture semantic nuances, as reflected in its lower scores for recall and nDCG. The introduction of BioBERT, leads to a significant boost across all metrics. This confirms the added value of semantic understanding, particularly crucial in specialized domains such as biomedicine, where simple lexical overlap between queries and documents is often insufficient.

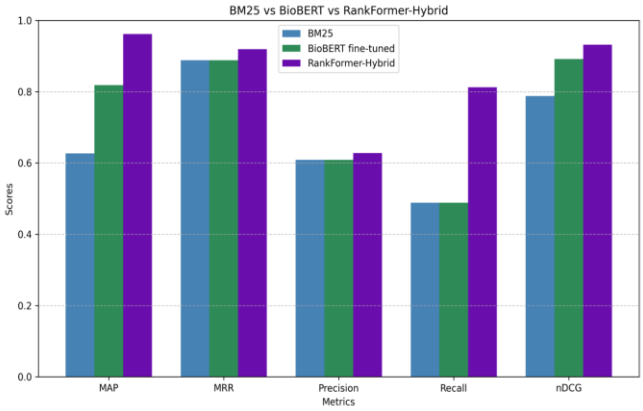


Fig. 3. Comparative performance of BM25, BioBERT, and the proposed RankFormer-Hybrid model.

The most substantial improvement is achieved by the RankFormer-Hybrid model, which combines lexical and semantic signals while leveraging supervised learning to refine document ranking. This model outperforms all others by a large margin, demonstrating the effectiveness of LTR approaches that not only understand content more deeply but also learn to prioritize relevance signals in a task-specific manner.

This comparative analysis underscores a key insight: while traditional and dense retrieval methods each contribute to performance, it is the combination of these methods within a supervised LTR framework that achieves the most robust and reliable results. In information retrieval tasks, especially those involving complex and high-stakes domains, incorporating LTR is not just beneficial but essential for achieving state-of-the-art performance.

To position our work within the current literature, we compare the performance of the proposed RankFormer-Hybrid model with that reported in selected previous studies. This qualitative comparison relies on standard evaluation metrics, including MAP, MRR, and nDCG.

Table VI presents a comparison between the proposed RankFormer-Hybrid model and other LTR models.

TABLE VI. COMPARISON OF THE PERFORMANCE OF THE PROPOSED RANKFORMER-HYBRID MODEL WITH OTHER LTR MODELS.

Model	Dataset	MAP@10	MRR@10	nDCG @10
BERT-based with injecting BM25 [27]	MS MARCO	0.367	0.364	0.422
HYRR [28]	MS MARCO	-	0.440	0.508
MultiLTR [29]	MQ2007	0.465	-	0.490
Vespa Hybrid [30]	TREC Covid	-	-	0.750
RankingRAE [31]	BioASQ Task B 2015	0.141	-	-
RankfFormer-Hybrid	BioASQ Task B 2024	0.961	0.920	0.932

The results in Table V show that the proposed RankFormer-Hybrid model outperforms several related models in terms of MAP, MRR, and nDCG at rank 10. Unlike methods that use BM25 scores as additional textual inputs or combine them through ensemble techniques, our approach directly integrates the hybrid BM25-BioBERT score into the document representation at the input level, allowing it to guide the attention mechanism from the earliest layers of the model. This integration leads to faster convergence, with the model reaching its best performance in only six epochs, which is particularly suitable for resource-constrained settings.

V. CONCLUSION

This work compares several information retrieval approaches on the BioASQ Task B 2024 dataset, ranging from the traditional BM25 (configured with k1 = 1.2, b = 0.8) to the semantic BioBERT model, and the hybrid learning-to-rank RankFormer approach. The results show a clear performance improvement as we move from sparse to dense and then to supervised hybrid methods, with RankFormer achieving the highest scores after just a few training epochs. These findings highlight the limitations of lexical-only retrieval in biomedical contexts and emphasize the effectiveness of incorporating semantic representations and hybrid signals. This study confirms the central role of learning-to-rank in modern IR systems and opens avenues for future research toward more adaptive and scalable ranking models.

A. Limitations and Discussions

While the results obtained in this study are encouraging and highlight the effectiveness of learning-to-rank techniques in biomedical information retrieval, some limitations should be acknowledged. First, the experiments are conducted exclusively on the BioASQ Task B 2024 dataset, which, although well established, may not fully represent the diversity and complexity of other biomedical or general domain IR tasks. As a result, the observed performance gains may not directly generalize to other biomedical retrieval tasks or to open-domain information retrieval settings without additional validation.

Second, although our methodology includes a hybrid scoring strategy and a deep LTR model, we limited our exploration to a single architecture and a relatively small number of training epochs. More comprehensive hyperparameter optimization and comparisons with alternative transformer-based ranking models could further strengthen the conclusions.

These limitations do not undermine the validity of the reported results; rather, they delineate the experimental boundaries of the present study and highlight directions for future research aimed at improving generalizability, robustness, and scalability.

B. Future Work

In future work, we plan to extend our evaluation to include other biomedical datasets to further assess the generalizability of our approach. Additionally, integrating user interaction signals could enable dynamic learning and continuous model improvement. Another promising direction involves exploring multi-model retrieval systems by incorporating structured knowledge or non-textual medical data into the ranking process. Finally, combining LTR with reinforcement learning or meta-learning techniques may offer new opportunities for optimizing

retrieval in highly specialized domains such as biomedical literature search.

REFERENCES

[1] H. Kulkarni, N. Goharian, O. Frieder, and S. MacAvaney, "LexBoost: improving lexical document retrieval with nearest neighbors," in ACM Symposium on Document Engineering, San Jose, CA, USA. ACM, New York, NY, USA, 10 pages, 2024.

[2] A. Goker and J. Davies, Information retrieval: searching in the 21st Century, ISBN: 9780470027622, John Wiley & Sons, Ltd, 2009.

[3] S. Sivarajkumar, H.A. Mohammad, D. Oniani, K. Roberts, W. Hersh, H. Liu, D. He, S. Visweswaran, and Y. Wang, "Clinical information retrieval: a literature review," Journal of Healthcare Informatics Research, 8(2):313–352, 2024.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Computation and Language, 2013.

[5] P.D. Turney and P. Pantel, "From frequency to meaning: vector space models of semantics," Journal of Artificial Intelligence Research, 37, 141-188, 2010.

[6] S.E. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Foundations and Trends in Information Retrieval, 3(4), 333-389, 2009.

[7] K.M. Svore and C.J.C. Burges, "A machine learning approach for improved BM25 retrieval," in Proceedings of the 18th ACM Conference on Information and Knowledge Management, HongKong, China, 2009.

[8] S.E. Robertson, H. Zaragoza, and M.J. Taylor, "Simple BM25 extension to multiple weighted fields," in Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, Washington, D.C., USA, 42-49, 2004.

[9] M. Luo, A. Mitra, T. Gokhale, and C. Baral, "Improving biomedical information retrieval with neural retrievers," Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

[10] P.S. Huang, X. He, J. Gao, et al., "Learning deep structured semantic models for web search using click through data," in Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, San Francisco, California, USA, 2333–2338, 2013.

[11] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 373–374, 2014.

[12] J. Devlin, M.W. Chang, K. Lee,and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186, 2019.

[13] T. Brown, B. Mann, N. Ryder, and D. Amodei, "Language models are few-shot learners," 34th Conference on Neural Information Processing Systems Vancouver,Canada, 2020.

[14] A.Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017.

[15] J. Lee, W. Yoon, Sungdong Kim, D. Kim , Sunkyu Kim, C.Ho So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, Volume 36, Issue 4, , Pages 1234–1240, 2020.

[16] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," ACM Transactions on Computing for Healthcare, 1(1), 24 pages, 2021.

[17] S. Bruch, S. Gai, and A. Ingber, "An analysis of fusion functions for hybrid retrieval," ACM Trans. Inf. Syst. 42, 1, 1-35, 2023.

[18] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender, "Learning to rank using gradient descent," in Proceedings of the 22nd international conference on Machine learning, New York, NY, USA, 89–96, 2005.

[19] Z. Cao, T. Qin, T.Y. Liu, M. F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in Proceedings of the 24th international conference on Machine learning, 129–136, 2007.

[20] M. Buyl, P. Missault, and P. A. Sondag, "RankFormer: listwise learning-to-rank using listwide labels," in Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 3762–3773, 2023.

[21] J. Lin, X. Ma, and A. Yates, "Pretrained transformers for text ranking: BERT and beyond," Synthesis Lectures on Human Language Technologies, 14(4), 1–325, 2021.

[22] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. Alvers, D. Weissenborn, A. Krithara, S.Petridis, S.A. Pavlopoulos, P.G. Bagos, J. Bjorne, S. Pyysalo, F. Ginter, T. Salakoski, M. Schroeder, E. Gaussier, M.R. Amini, Y. Kouropalatis, C. Matuschek, L. Goeuriot, E. Gaussier, M. Stevenson, S. Ananiadou, A.Kastrin, and M. Dehmer. "An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition," BMC Bioinformatics, 16 :138, 2015.

[23] J. Lin, X. Ma, S.C. Lin, J. H. Yang, R. Pradeep, and R. Nogueira, "Pyserini: an easy-to-use Python toolkit to support replicable IR research with sparse and dense representations," in proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021.

[24] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and E. M. Voorhees, "Overview of the TREC 2019 deep learning track," in TREC 2019 Conference Notebook. National Institute of Standards and Technology, 2020.

[25] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, ISBN: 0262035618, 2016.

[26] S. Chavhan, M. Raghuwanshi, and R.C. Dharmik, "Information retrieval using machine learning for ranking: a review," Journal of Physics Conference Series 1913(1):012150, 2021.

[27] A. Askari, A. Amin, P. Gabrielle, K. Wessel, and V. Suzan, "Injecting the BM25 score as text improves BERT-based re-rankers," in proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), pages 1682–1686, 2023.

[28] J. Lu, H. Keith, B. Oguz, M. Ji, and N. Jianmo, "HYRR: Hybrid infused reranking for passage retrieval," in proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 139–151, 2022.

[29] H. Yang and T. Gonçalves, ""MultiLTR: Text ranking with a multi-stage learning-to-rank approach," Information, 16(4) :308, 2025.

[30] J. K. Bergum, "Improving zero-shot ranking with vespa hybrid search – part two," https://blog.vespa.ai/improving-zero-shot-ranking-with-vespa-part-two/, 2023. Vespa Blog.

[31] Y. Yan, B. Zhang, and L. Xu-Feng, "RankingRAE : Global learning to rank for question answering with recursive autoencoders," PLOS ONE, 15(11) :e0242061, 2020.

Optimizing Rainfall Prediction in Settat, Morocco, Through Machine Learning

Oussama Zemnazi[1], Sanaa El Filali[2], Sara Ouahabi[3], Abderrahim Mouhtadi[4]
Faculty of Sciences Ben M'Sick-Laboratory of Artificial Intelligence and Systems,
University Hassan II - Casablanca, Casablanca, Morocco[1, 2, 3]
General Directorate of Meteorology, Morocco[4]

***Abstract*—Rainfall prediction is still a difficult challenge because rainfall is nonlinear, intermittent, and highly variable, especially in semi-arid climates. Accurate rainfall prediction is crucial for water resource management, agricultural planning, climate-driven decision-making, and more. This study proposes a comparative framework based on machine learning and ensemble learning techniques to predict daily rainfall in Settat, Morocco, as a representative semi-arid region. Five predictive models were trained and evaluated based on meteorological station observations: Random Forest, XGBoost, LightGBM, CatBoost, and a Multilayer Perceptron (MLP). The models' performance was evaluated based on mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE), and the coefficient of determination (R-squared). The results demonstrate that the performance and stability of gradient boosting algorithms are superior to all other evaluated models. Specifically, LightGBM produced the fewest erroneous values and explained rainfall variability best. These results underscore the success of boosting-based ensemble techniques in modeling inconsistent precipitation patterns and provide a comparative framework for machine-learning-based rainfall forecasting in semi-arid environments.**

***Keywords*—Rainfall forecasting; machine learning; gradient boosting; LightGBM; semi-arid climate; ensemble learning**

I. Introduction

The forecast of rainfall is vital in water resource management, agricultural productivity, flood risk intervention, and climate resilience strategies, especially in semi-arid environments where precipitation is rare, erratic, and variable. In these regions, rainfall tends to have long dry spells and short, intense storms, leading to uncertain water resource management and a high susceptibility to droughts and flash floods [1], [2].

Traditionally, rainfall prediction has relied on Numerical Weather Prediction (NWP), which is based on the numerical integration of atmospheric physical equations. Although these models provide accurate predictions at synoptic and regional scales, they perform poorly at the local and station levels, especially in data-limited semi-arid areas [3]. Furthermore, NWP models have high computational costs and are driven by initial conditions; they generally do not capture orographic rainfall processes that trigger convection and dominate precipitation variability in dry climates [4].

Recently, machine learning (ML) methods have been established as competitive alternatives for rainfall prediction due to their ability to learn complex nonlinear relationships directly from historical meteorological data. Ensemble learning methods, such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), LightGBM, and CatBoost, are widely used for tabular climate data because of their robustness, ability to handle heterogeneous predictors, and reduced sensitivity to overfitting [5]–[8]. Neural network models, especially the multilayer perceptron (MLP), have been extensively used for rainfall forecasting due to their ability to model complex nonlinear relationships. These models perform effectively when the feature space is well engineered and when large amounts of training data are available, which can undermine their performance in the context of station-based semi-arid areas [9].

Even with the growing use of ML techniques for rainfall prediction, several limitations remain. First, most previous work focuses on temperate or data-rich regions. However, semi-arid settings are under-investigated despite rainfall intermittency, skewed distributions, and heavy-tailed distributions (extreme events) [2], [10]. Second, existing works commonly employ heterogeneous preprocessing techniques, feature representations, and evaluation metrics, which make comparisons among models overly challenging. Consequently, there is still no clear consensus on whether ensemble boosting algorithms consistently outperform neural network models for daily rainfall prediction when evaluated under identical experimental conditions. In addition, relatively few studies provide climate-aware interpretations that explicitly link model performance to the physical characteristics of rainfall processes in semi-arid regions [11].

To bridge the gap, a comprehensive and systematic comparison of five machine learning models for daily rainfall prediction is conducted in Settat, Morocco, a representative semi-arid area. The compared models are four ensemble models: LightGBM, XGBoost, CatBoost, and the baseline neural Multilayer Perceptron (MLP). All models are trained and evaluated with the same 11 years of daily meteorological observation data using the same processing, feature engineering, and evaluation metrics. This unified framework enables a fair and transparent assessment of the relative strengths and limitations of each modeling approach.

The main contributions of this study are summarized as follows:

- A universal benchmarking system to evaluate the ensemble and neural network methods in a daily rainfall forecasting task with identical experimental settings.

- The analysis is based on long-term station observations of a semi-arid climate with high rainfall intermittency.
- A climate-informed assessment of model performance, focusing on the applicability of gradient boosting techniques to tabular meteorological data in data-sparse settings.

The rest of the study is organized as follows: Section II presents related work on machine learning-based rainfall prediction. Section III provides the area of study, the dataset, and the preprocessing. The modeling techniques and evaluation metrics are described in Section IV. Section V presents the experimental results, and we discuss our findings and their limitations in Section VI. Section VII concludes the study and outlines future research directions.

II. Related Work

A. Rainfall Forecasting Using Machine Learning

The prediction of rainfall has long been acknowledged as a difficult problem, primarily because precipitation processes are known to be both nonlinear and nonstationary with high intermittency. Traditional statistical methods, including linear regression and autoregressive models, have been widely used for rainfall prediction. However, their predictive performance is typically weak under strong climate fluctuations and extreme rainfall events. These constraints have led to the growing use of machine learning (ML) approaches as data-driven solutions for rainfall prediction.

Machine learning techniques have demonstrated strong capabilities for capturing complex nonlinear associations between rainfall and atmospheric predictors. Various ML techniques, such as decision trees, support vector machines, artificial neural networks, and ensemble learning methods, have been used to predict rainfall using station-based gauges, satellite-derived products, or reanalysis data. Several researchers have found that machine learning (ML) methods can outperform traditional statistical approaches for daily rainfall forecasting, especially when multiple meteorological variables are used as inputs [2], [12]. Reanalysis data, including ERA5, have also been increasingly adopted in ML-based rainfall prediction to provide spatially and temporally consistent atmospheric information [10]. Nevertheless, the performance of ML techniques on rainfall prediction is still sensitive to regional climate conditions, data availability, and experimental settings. Most previous research has been carried out in wet or temperate zones, since rainfall patterns there are relatively regular. In contrast, semi-arid areas (with low, unevenly distributed rainfall) have received less attention [2]. This disparity underscores the need for region-specific studies that explicitly address the challenges of rainfall prediction in semi-arid environments.

B. Boosting and Neural Network Models for Rainfall Prediction

Ensemble-based methods and neural networks are the two main modeling paradigms for rainfall forecasting among machine learning techniques. Tree-based ensemble learners, including Random Forest, have been extensively used in rainfall prediction because of their strength in modeling when interactions exist among input predictors and their efficacy in handling the heterogeneity of meteorological covariates [5]. Random Forest models have been shown to perform well in daily rainfall prediction and can yield better results than classical regression methods, particularly in systems with high rainfall variability [12], [13].

Gradient boosting methods such as XGBoost and LightGBM have received growing attention in recent rainfall prediction research. These models refine prediction accuracy by correcting residual errors. They are found to perform significantly better than single decision trees and classical neural networks in several rainfall estimation studies [6], [14], [15]. Their excellent ability to handle small sample sizes, missing values, and skewed rainfall distributions makes them particularly desirable for station-based rainfall forecasting in data-limited areas. CatBoost has also been investigated for precipitation prediction, with reported improvements in robustness and reduced bias in simulating complex rainfall patterns [8], [16].

Neural network techniques have also been widely studied in the context of rainfall prediction. Multilayer Perceptron (MLP) networks have been employed for daily rainfall forecasting using data from meteorological stations. They can produce satisfactory results when large volumes of training data are used and careful input feature selection is performed [9], [17]. More complex neural network structures have also been investigated to capture temporal relationships in precipitation series. However, these methods usually require substantial data and extensive hyperparameter tuning, making them challenging to use in station-based and semi-arid contexts.

Comparison analysis shows that boosting-based ensemble models generally yield more stable and robust predictions than neural network methods for tabular meteorological data and daily rainfall forecasting, especially in the presence of rainfall intermittency and scarce data [14], [15]. However, model performance is heavily dependent on climate, predictor selection, and evaluation methods, underscoring the need for consolidated benchmarking.

C. Research Gaps and Challenges in Semi-Arid Regions

Although machine learning models have attracted attention in rainfall prediction, there is still little research on this topic, especially in semi-arid areas. First, many studies were conducted under heterogeneous conditions for preprocessing, feature engineering, and evaluation metrics. This lack of uniformity prevents comparisons of reported results and makes it difficult to draw strong conclusions about model performance. Second, in semi-arid climates, rainfall occurs as long dry spells punctuated by brief, heavy rainfall events. This type of sparsity, intermittency, and skewness can pose major challenges in the development and performance of data-driven models when identifying hidden patterns from scarce rainfall observations. Therefore, performance observed in humid or data-rich regions cannot be extrapolated to semi-arid conditions without rigorous validation. Finally, few studies offer climate-sensitive interpretations of the performance of machine learning models directly related to the physical characteristics of rainfall processes in arid areas. Existing work tends to be accuracy-oriented and pays little attention to what makes the model perform better or worse in a particular climate. Addressing these

challenges requires unified experimental frameworks, region-specific analysis, and interpretative discussions that link model behavior to rainfall variability and data characteristics in semi-arid regions.

III. Study Area and Data Description

A. Study Area

The research was carried out in Settat, a city situated in west-central Morocco, and covers the semi-arid climatic zone. Settat is located at altitudes ranging from 200 to 416 m above sea level and is characterized by high interannual and intra-annual rainfall variability. The area is known for a regional weather regime characterized by long, dry seasons terminated by short, sporadic rainfall events, making precipitation prediction particularly difficult.

Seasonal temperature variations are pronounced. Maximum air temperatures in the summer period commonly exceed 35°C and can reach up to 46°C, while those in winter are generally milder (5 to 15°C). Precipitation is low throughout the year and varies from one month to another. These conditions render Settat a suitable study area for assessing the reliability of machine learning techniques for precipitation estimation in semi-arid climates, where precipitation processes are highly intermittent and nonlinear.

B. Data Sources and Description

The meteorological data used in this work are obtained from the Settat weather station of the General Directorate of Meteorology (GDM), Morocco. The dataset comprises daily records for 2014-2024. Owing to institutional policy on data sharing, raw data will not be made publicly accessible unless an individual requests it with the GDM.

The dataset consists of eight daily meteorological variables selected for their potential influence on rainfall generation processes. These variables include maximum air temperature (TMPMAX, °C), minimum air temperature (TMPMIN, °C), mean station pressure (Mean_Pressure, hPa), maximum relative humidity (HUMAXQ, %), minimum relative humidity (HUMINQ, %), maximum wind speed (FFXINS, m/s), maximum wind direction (DDXINS, °), and daily rainfall (RRQUOT, mm). Daily rainfall (RRQUOT) was used as the target variable, with all other weather variables as predictors. The selection of these variables is guided by their physical importance to the process and by their accessibility at the station scale.

C. Data Preprocessing

A uniform and well-designed preprocessing workflow was adopted for the dataset, maintaining its temporal structure and improving the model's learning. In a first step, meteorological data were ordered chronologically to match the time scales. In cases of missing values, we linearly interpolated them to produce a continuous time series while avoiding the introduction of artificial discontinuities. Cyclical encoding of the calendar variables was used to model the seasonal and temporal patterns. Precisely, the day and month information was processed by sine and cosine functions, and binary dummy variables were included to denote the four meteorological seasons. Furthermore, lagged features at several past time points were derived from the original predictors to capture temporal dynamics. This was in addition to rolling statistical features (daily and cumulative means, maxima, minima, standard deviations, and daily differences), which were also calculated for persistence and short-term variability in meteorological conditions.

Since tree-based ensemble models (Random Forest, XGBoost, LightGBM, and CatBoost) are not sensitive to feature scaling, we used the generated features directly. On the other hand, we get that the MLP model requires feature scaling; therefore, Min-Max normalization was applied only after the chronological data split to avoid data leakage. The data was divided into 70% for training, 15% for validation, and 15% for testing in a strict temporal order. Finally, non-informative attributes such as station IDs and raw date columns were dropped from the dataset, leaving only those with significance for weather conditions or the period. Linear interpolation was selected due to the relatively low proportion of missing values and its ability to preserve the temporal continuity of daily meteorological time series without introducing artificial variability.

It is important to note that this study is based on data from a single meteorological station located in Settat. This choice is primarily due to the limited availability of high-quality and continuous multi-station meteorological data in the region.

However, the Settat station is representative of the semi-arid climatic conditions of central Morocco, characterized by low, irregular rainfall. Therefore, the findings of this study provide meaningful insights into rainfall prediction in similar semi-arid environments.

Future work will focus on extending this approach to multi-station datasets and incorporating spatial variability to improve model generalization further.

IV. Methodology

A. Experimental Framework

In this research, a consolidated, organized experimental framework is proposed to enable an equitable comparison of machine learning models for daily rainfall prediction. The purpose of this study was to compare the predictive performance of ensemble tree models and neural network approaches under the same data requirements and evaluation protocols. All models were trained using the feature sets generated by the preprocessing pipeline described in Section III. The datasets were divided into 70% for training, 15% for validation, and 15% for testing while preserving the time sequences (no data leakage). Hyperparameters were optimized for the training and validation subsets only, and the test set was held off until the final performance assessment. This controlled setup ensures both methodological comparability and statistical reliability when comparing models. The overall workflow of the proposed methodology is illustrated in Fig. 1.

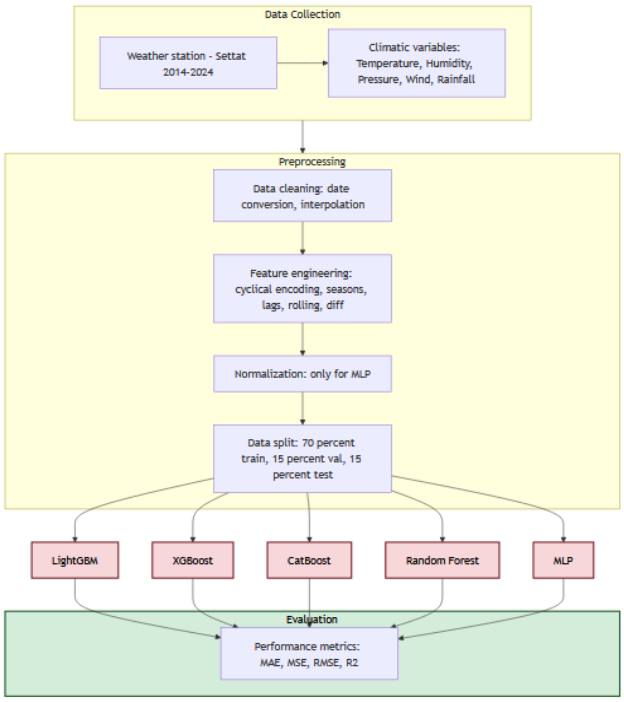


Fig. 1. Proposed methodology flowchart.

B. Machine Learning Models

The five supervised regression models used in this study were Random Forest, XGBoost, LightGBM, CatBoost, and Multilayer Perceptron (MLP). These models were selected to represent two dominant paradigms in tabular data modeling: ensemble-based tree methods and neural network architectures.

1) Random Forest (RF): Random Forest is an ensemble learning method that creates multiple decision trees via bootstrap sampling and random feature selection. The final prediction is obtained by averaging the three results. This strongly reduces variance and enhances generalization. Random forest is well-suited for a precipitation forecast model in a semi-arid region with an irregular and highly variable rainfall distribution because of its noise insensitivity and ability to handle multicollinearity and non-linear interactions.

2) Extreme Gradient Boosting (XGBoost): XGBoost is a gradient-boosting machine learning method that trains decision trees sequentially to minimize a regularized objective. The model adjusts the residuals from the previous trees at each iteration. The model also includes regularization to prevent overfitting and ensure the stability of the training model. XGBoost is well-known for its excellent predictive performance on structured data with nonlinear patterns.

3) Light Gradient Boosting Machine (LightGBM): LightGBM is a gradient boosting framework that uses a histogram-based learning algorithm, and it grows the tree leaf-wise. These models improve computational efficiency while preserving high predictive performance. LightGBM is particularly suited to structured meteorology datasets with engineered statistical and temporal features.

4) CatBoost: CatBoost is a gradient boosting algorithm that aims to reduce prediction bias by using ordered boosting methods. Although the data source used in this study contains mostly numerical features, we included CatBoost because it has been reported to be robust to noise and performs well for general tabular regression problems.

5) Multilayer Perceptron (MLP): The Multilayer Perceptron is a feedforward artificial neural network consisting of fully connected layers with nonlinear activation functions. For this analysis, an MLP was used as a regression model to capture complex nonlinear relationships between meteorological variables and daily rainfall. Neural Networks are sensitive to variations in feature scale. Therefore, the input variables were normalised using min–max normalisation before model training.

C. Hyperparameter Optimization

Hyperparameter optimization was performed on the training and validation subsets. Ensemble model hyperparameters, including the number of trees, maximum tree depth, and learning rate, were optimized to maximize predictive performance without overfitting. For the MLP network, the hidden neurons, learning rate, and number of training iterations were tuned. The final hyperparameter settings for each model were chosen based on validation performance. The test set was held out entirely during model tuning to obtain an unbiased estimate. The final selected hyperparameter values for each model are summarized in Table I.

TABLE I. HYPERPARAMETER SETTINGS OF THE EVALUATED MODELS

Model	Hyperparameters
LightGBM	learning_rate = 0.03; num_leaves = 31; feature_fraction = 0.8; bagging_fraction = 0.8; min_data_in_leaf = 20
XGBoost	learning_rate = 0.03; max_depth = 3; subsample = 0.8; colsample_bytree = 0.8; min_child_weight = 2
CatBoost	iterations = 1000; learning_rate = 0.03; depth = 6; l2_leaf_reg = 3
Random Forest	n_estimators = 200; max_depth = None
MLP	hidden_layers = (64, 32); activation = ReLU; learning_rate = 0.001; batch_size = 32; epochs = 120

Early stopping was applied for gradient boosting models and the MLP model to determine the optimal number of training iterations and prevent overfitting.

D. Evaluation Metrics

Model performance was evaluated using four complementary regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). The MAE is the mean of the differences between observed and predicted rainfall values. MSE gives the mean of the squared errors and is sensitive to large deviations. RMSE: The square root of the MSE, RMSE, quantifies rainfall prediction error (millimetres), enabling interpretation. R^2 is the fraction of the variability in observed rainfall that the model explains. Their joint consideration enables accurate measurement of the quality and performance of models in semi-arid regions with low precipitation.

V. RESULTS

Before evaluating model performance, it is important to analyze the statistical characteristics of the rainfall data. The dataset is characterized by a high proportion of zero rainfall values, corresponding to long dry periods typical of semi-arid climates. In contrast, rainfall events occur sporadically and are often characterized by high intensity, resulting in a highly skewed distribution.

This intermittency and variability make rainfall prediction particularly challenging, as models must simultaneously capture both the absence of rainfall and sudden extreme events. These characteristics highlight the importance of using robust machine learning models that can handle nonlinear and imbalanced data distributions.

A. Overall Predictive Performance

The performance of the tested machine learning models was evaluated on the test set using four evaluation metrics: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R^2). A comparative result for Random Forest (RF), XGBoost, LightGBM, CatBoost, and Multilayer Perceptron (MLP) is shown in Table II.

TABLE II. PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS FOR DAILY RAINFALL PREDICTION.

Metrics	Method				
	LightGBM	***XGBoost***	***CatBoost***	***RF***	***MLP***
MAE	0.0507	0.0423	0.0534	0.0683	0.1628
MSE	0.0456	0.0530	0.0536	0.1199	0.3096
RMSE	0.2135	0.2301	0.2314	0.3463	0.5564
R^2	0.9868	0.9846	0.9844	0.9652	0.9107

The performance of the evaluated machine learning models is summarized in Table II using four evaluation metrics: MAE, MSE, RMSE, and R^2. The results clearly indicate that gradient boosting models outperform the other approaches across all metrics.

LightGBM achieved the best overall performance, with the lowest error values and the highest coefficient of determination ($R^2 = 0.9868$), indicating its strong ability to accurately model rainfall variability. XGBoost and CatBoost also demonstrated competitive performance, with slightly higher error values but still maintaining high predictive accuracy.

In contrast, the Random Forest model showed moderate performance, with higher error rates than boosting methods. The MLP model exhibited the weakest performance, with significantly higher errors and lower R^2 values, indicating difficulties in learning from intermittent, highly variable rainfall patterns.

The superior performance of gradient boosting models can be attributed to their sequential learning strategy, which iteratively reduces prediction errors and enhances model generalization. These models are particularly effective in handling nonlinear relationships and skewed data distributions, which are common characteristics of rainfall in semi-arid regions.

Furthermore, boosting algorithms are more robust to sparse data and are better suited to capture rare but significant rainfall events. In contrast, neural network models such as MLPs require larger, more balanced datasets to achieve stable performance, which limits their effectiveness in this context.

B. Feature Importance Analysis

To further interpret the predictive behavior of the best-performing model, a feature importance analysis was conducted using the LightGBM algorithm. This analysis provides insight into the relative contribution of each input variable to the rainfall prediction task.

The results indicate that meteorological variables such as relative humidity, temperature, and wind speed play a significant role in rainfall prediction. In particular, relative humidity was identified as one of the most influential features, as it directly reflects atmospheric moisture conditions necessary for precipitation formation.

Temperature variables also contribute significantly, as they influence evaporation and condensation processes. Wind-related features further enhance prediction performance by capturing atmospheric dynamics and moisture transport.

Overall, the feature importance analysis confirms that the selected input variables are physically relevant and consistent with the underlying climatic processes governing rainfall in semi-arid regions. This strengthens the reliability and interpretability of the proposed machine learning approach.

C. Observed vs. Predicted Daily Rainfall

To better understand the predictive behavior of the evaluated models, a visual comparison between observed and predicted daily rainfall values is presented for selected models. The selected models include LightGBM as the best-performing model, XGBoost as a competitive gradient boosting approach, and the Multilayer Perceptron (MLP) as a neural network baseline.

Fig. 2 illustrates the comparison between observed and predicted daily rainfall using the LightGBM model on the testing dataset. The predicted rainfall closely follows the observed temporal patterns, effectively capturing both prolonged dry periods and rainfall events. This demonstrates the model's strong capability in representing the nonlinear and intermittent nature of rainfall in semi-arid regions. However, slight underestimation is observed during high-intensity rainfall events, which is consistent with the difficulty of modeling extreme values in skewed precipitation distributions.

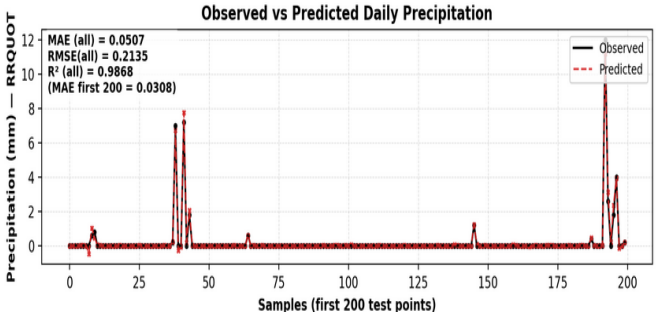


Fig. 2. Observed vs. predicted daily rainfall using LightGBM.

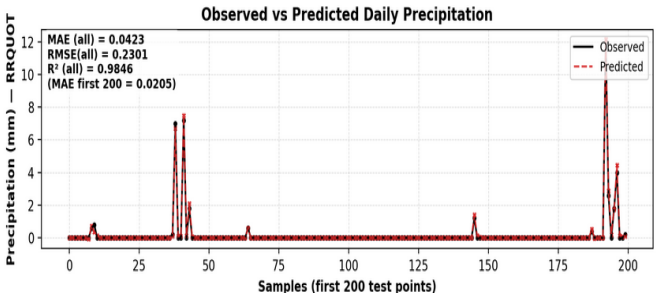


Fig. 3. Observed vs. predicted daily rainfall using XGBoost.

Fig. 3 presents the observed and predicted rainfall values using the XGBoost model. Similar to LightGBM, XGBoost can capture general rainfall dynamics and temporal variability with high accuracy. Nevertheless, minor deviations are evident compared to LightGBM, particularly during peak rainfall periods, suggesting slightly lower precision in modeling extreme events.

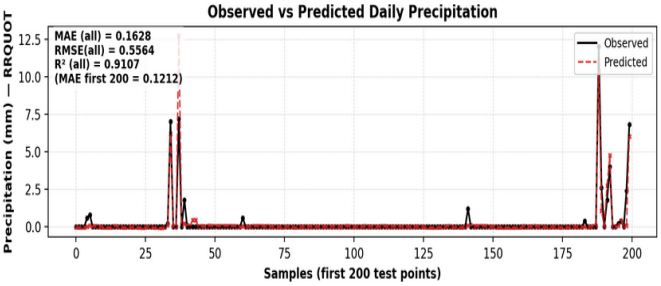


Fig. 4. Observed vs. predicted daily rainfall using MLP.

Fig. 4 shows the MLP model's performance in predicting daily rainfall. Unlike the boosting-based methods, the MLP model exhibits larger discrepancies between observed and predicted values. The model struggles to accurately reproduce rainfall variability, especially during abrupt transitions and extreme rainfall events. This behavior reflects the limitations of neural networks when applied to relatively small and highly intermittent meteorological datasets.

Overall, the visual comparison confirms the quantitative results presented earlier, demonstrating that gradient boosting models, particularly LightGBM and XGBoost, provide more accurate and stable predictions than the neural network model. These models are better suited to capture the intermittent, nonlinear, and skewed characteristics of rainfall in semi-arid environments.

D. Comparative Analysis of Model Behavior

Model comparison reveals the robustness and stability of gradient boosting models (LightGBM, XGBoost, and CatBoost) across rainfall intensity levels.

LightGBM has the lowest MAE, MSE, and RMSE values and the highest R^2, indicating a stronger ability to minimize prediction error while explaining more rainfall variability.

In contrast, the MLP model displays greater variability in prediction errors at all-time steps, especially during rapid precipitation changes. Random forests show reasonable robustness but are still behind boosting-based methods on the respective metrics evaluated.

These results indicate that boosting algorithms are particularly ideal for simulating the intermittent and skewed rainfall distributions commonly found in semi-arid environments.

The graphical comparison between observed and predicted rainfall values further supports these findings. As illustrated in Fig. 2 to Fig. 4, the boosting-based models provide a closer fit to the observed data, whereas the MLP model exhibits larger deviations, particularly during abrupt rainfall changes.

VI. Discussion

A. Interpretation of Model Performance

The experimental results indicated that all gradient boosting models, especially LightGBM, performed strongly in reducing prediction errors and explaining rainfall variability when predicting daily rainfall in the semi-arid region of Settat. For instance, a lower MSE value implies greater reliability against large prediction errors, and lower MAE and RMSE suggest improved consistency in reproducing rainfall intensity. The higher R^2 score also suggests that this model better explains rainfall variability than those analyzed.

Boosting algorithms are particularly strong because they iteratively refine residual errors to improve generalization via a sequential error-correction mechanism. This framework is especially beneficial for modelling nonlinear relationships between meteorological predictors and for accommodating skewed rainfall distributions. There have been similar findings in semi-arid rainfall prediction studies using ensemble and boosting methods [18], [19]. In contrast, broader reviews stress that AI-based rainfall forecasting frameworks are highly robust to climatic variability [20].

Conversely, the MLP showed greater variation in prediction errors, especially during rapid rainfall changes. Neural network models need more extensive, balanced datasets to achieve stable performance. Similar findings from comparative studies across various climatic regions indicate that deep learning models may struggle when trained on limited or highly intermittent rainfall datasets [21].

B. Comparison with Existing Literature

To evaluate the effectiveness of the proposed approach, the obtained results were compared with findings from recent studies on rainfall prediction in semi-arid regions.

Recent studies have demonstrated the effectiveness of machine learning and ensemble techniques for precipitation modeling. For instance, a recent study conducted in Morocco [18] reported high predictive performance using CatBoost and XGBoost models, with R^2 values reaching approximately 0.98. Similarly, other studies have demonstrated that machine learning and hybrid approaches achieve strong predictive performance in rainfall prediction tasks across different climatic conditions [22], [23].

In comparison, the proposed LightGBM model achieved an R^2 of 0.9868, demonstrating competitive, robust predictive performance. These results confirm that ensemble learning approaches are well-suited for modeling the nonlinear and highly variable nature of rainfall in semi-arid environments.

C. Implications and Study Limitations

Semi-arid areas are characterized by long dry periods interspersed with short-duration, high-intensity events, making rainfall forecasting a fundamental challenge in these regions. This intermittency results in highly skewed data distributions, making models sensitive to extreme values. The reduction in RMSE for the LightGBM across both extremes was observed, confirming its resilience to these extremes and further suggesting a move towards semi-arid rainfall modelling with this algorithm.

However, some limitations need to be considered. First, the analysis is based on a single meteorological station, which limits spatial generalizability. Second, daily one-step-ahead forecasting was only examined when multiple-step-ahead prediction frameworks may yield more operationally relevant insights [24]. Third, the more sophisticated spatiotemporal deep learning models were not applied in this study [25].

Nevertheless, the general, unified evaluation framework used in this study provides a clear standard against which to compare machine learning algorithms in semi-arid climates. It contributes to understanding the suitability of algorithms under intermittent precipitation regimes.

VII. Conclusion

This study presented a comprehensive evaluation of machine learning models for daily rainfall prediction in a semi-arid region of Morocco. Several models, including LightGBM, XGBoost, Random Forest, CatBoost, and MLP, were assessed using standard performance metrics.

The results demonstrated that the LightGBM model achieved the best performance, with a high coefficient of determination (R^2 = 0.9868) and low prediction errors, indicating its strong capability to model nonlinear and highly variable rainfall patterns.

These findings highlight the effectiveness of ensemble learning approaches for rainfall prediction, particularly in data-scarce and semi-arid environments.

However, this study is limited by the use of data from a single meteorological station, which may affect the generalization of the results.

Future work will focus on extending this approach to multi-station datasets, incorporating spatial and temporal variability, and exploring advanced deep learning architectures, such as LSTMs and hybrid models, to further improve prediction accuracy.

Acknowledgment

The authors would like to thank the General Directorate of Meteorology (DGM) in Morocco for providing the meteorological data used in this study.

References

[1] M. V. K. Sivakumar, A. C. Ruane, and J. Camacho, "Climate change in the West Asia and North Africa region," in Climate Change and Food Security in West Asia and North Africa, Dordrecht, The Netherlands: Springer, 2013, pp. 3–26, doi: 10.1007/978-94-007-6751-5_1.

[2] L. V. Alexander et al., "Global observed changes in daily climate extremes of temperature and precipitation," J. Geophys. Res. Atmos., vol. 111, no. D5, Art. No. D05109, 2006, doi: 10.1029/2005JD006290.

[3] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," Nature, vol. 525, no. 7567, pp. 47–55, 2015, doi: 10.1038/nature14956.

[4] M. G. Schultz et al., "Can deep learning beat numerical weather prediction?" Phil. Trans. R. Soc. A, vol. 379, no. 2194, Art. no. 20200097, 2021, doi: 10.1098/rsta.2020.0097.

[5] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discover. Data Min. (KDD), San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[7] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 3146–3154.

[8] L. Prokhorenkova et al., "CatBoost: Unbiased boosting with categorical features," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Montréal, QC, Canada, 2018, pp. 6638–6648.

[9] A. Maier and A. Dandy, "Neural networks for the prediction and forecasting of water resources variables: A review," Environmental Modelling & Software, vol. 15, no. 1, pp. 101–124, 2000, doi: 10.1016/S1364-8152(99)00007-9.

[10] H. Hersbach et al., "The ERA5 global reanalysis," Q. J. R. Meteorol. Soc., vol. 146, no. 730, pp. 1999–2049, 2020, doi: 10.1002/qj.3803.

[11] M. Reichstein et al., "Deep learning and process understanding for data-driven Earth system science," Nature, vol. 566, no. 7743, pp. 195–204, 2019, doi: 10.1038/s41586-019-0912-1.

[12] C. M. Liyew and H. A. Melese, "Machine learning techniques to predict daily rainfall amount," Journal of Big Data, vol. 8, no. 153, 2021, doi: 10.1186/s40537-021-00545-4.

[13] M. Abu Saleh, H. M. Rasel, and B. Ray, "A comprehensive review towards resilient rainfall forecasting models using artificial intelligence techniques," Green Technologies and Sustainability, vol. 2, No. 3, Art. no. 100104, 2024, doi: 10.1016/j.grets.2024.100104.

[14] A. Sreenivasu, S. Rafi, and C. Rajani, "Rainfall prediction using machine learning," in 2024 IEEE International Conference on Machine Learning and Computing (ICMACC), 2024, pp. 86–90, doi: 10.1109/ICMACC62921.2024.10894486.

[15] V. Kumar, N. Kedam, O. Kisi, S. Alsulamy, K. M. Khedher, and M. A. Salem, "A comparative study of machine learning models for daily and weekly rainfall forecasting," Water Resources Management, vol. 39, no. 1, pp. 271–290, 2025, doi: 10.1007/s11269-024-03969-8.

[16] V. Kumar, N. Kedam, K. V. Sharma, K. M. Khedher, and A. E. Alluqmani, "A comparison of machine learning models for predicting rainfall in urban metropolitan cities," Sustainability, vol. 15, no. 18, Art. no. 13724, 2023, doi: 10.3390/su151813724.

[17] I. V. Necesito, D. Kim, Y. H. Bae, K. Kim, S. Kim, and H. S. Kim, "Deep learning-based univariate prediction of daily rainfall: application to a flood-prone, data-deficient country," Atmosphere, vol. 14, no. 4, Art. no. 632, 2023, doi: 10.3390/atmos14040632.

[18] A. Elmotawakkil, A. Moumane, A. Sadiki, et al., "Forecasting short-term rainfall patterns in arid and semi-arid regions using machine learning and deep learning models: A case study from Morocco," Theoretical and Applied Climatology, 2025, doi: 10.1007/s00704-025-05677-8.

[19] M. El Hafyani, K. El Himdi, and S. E. El Adlouni, "Improving monthly precipitation prediction accuracy using machine learning models: A multi-view stacking learning technique," Frontiers in Water, vol. 6, 2024, doi: 10.3389/frwa.2024.1378598.

[20] F. A. F. Sham, A. El-Shafie, W. Z. Wan Jaafar, M. Sherif, and A. N. Ahmed, "Advances in AI-based rainfall forecasting: A comprehensive review of past, present, and future directions with intelligent data fusion and climate change models," Results in Engineering, vol. 27, 2025, doi: 10.1016/j.rineng.2025.105774.

[21] O. A. Wani et al., "Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the north-western Himalayas," Scientific Reports, vol. 14, 2024, doi: 10.1038/s41598-024-77687-x.

[22] A. K. Sharma, R. Gupta, and S. Verma, "How accurate are machine learning models in improving monthly rainfall prediction in hyper-arid environments?," Journal of Hydrology, vol. 633, 2024, doi: 10.1016/j.jhydrol.2024.131040.

[23] M.-H. Lee and Y. J. Chen, "Precipitation modeling for extreme weather based on sparse hybrid machine learning and Markov chain random field," Water, vol. 13, no. 9, p. 1241, 2021, doi: 10.3390/w13091241.

[24] S. Narejo et al., "Multi-step rainfall forecasting using deep learning approach," PeerJ Computer Science, vol. 7, e514, 2021, doi: 10.7717/peerj-cs.514.

[25] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," arXiv preprint arXiv:1706.03458, 2017.

A Systematic Literature Review on Artificial Intelligence Applications for Breast Cancer Classification

Nursakinah Abdullah[1], Qi Wei Oung[2], Chee Chin Lim[3],
Chiew Chea Lau[4], Vrshni Menaka R Siva Nathan[5], Hui Wen Tiu[6]
Faculty of Electronic Engineering & Technology (FKTEN), Universiti Malaysia Perlis, Arau, 02600, Perlis, Malaysia [1, 2, 3, 6]
Sport Engineering Research Centre (SERC), Universiti Malaysia Perlis, Arau, 02600, Perlis, Malaysia[3]
Advanced Communication Engineering (ACE), Universiti Malaysia Perlis, Arau, 02600, Perlis, Malaysia[2]
School of Medical Sciences-Health Campus, Universiti Sains Malaysia,
Kubang Kerian, Kota Bharu, 16150, Kelantan, Malaysia[4, 5]

***Abstract*—Breast cancer remains one of the most prevalent and life-threatening diseases worldwide, needing to be diagnosed early and properly classified for effective treatment. Advancements in artificial intelligence (AI), deep learning, and machine learning techniques have shown great potential in automating breast cancer diagnosis and molecular subtyping using medical imaging. This systematic literature review explores the application of AI in breast cancer classification, focusing on mammographic imaging and its application in distinguishing molecular subtypes. The study follows the PRISMA guideline, investigating studies from multiple digital libraries published between 2020 and November 2024. Findings show that while deep learning models have significantly improved breast cancer detection, challenges remain in optimizing classification models for molecular subtypes, balancing accuracy and interpretability, and integrating AI-based tools into clinical practice workflows. Besides, heterogeneity in preprocessing pipeline algorithms and dataset limitations highlights the importance of conducting additional research to develop robust and generalized classification models. This review underscores the importance of AI-driven solutions in advancing breast cancer diagnosis and treatment planning while providing insights into future research directions.**

Keywords*—Artificial intelligence; breast cancer; classification; Convolutional Neural Network; deep learning; machine learning; mammography; medical imaging; molecular subtypes; Vision Transformer

I. Introduction

Breast cancer occurs when abnormal cells in the breast grow uncontrollably, forming tumors. This often involves changes in hormone receptors, such as estrogen (ER), progesterone (PR), and human epidermal growth factor receptor 2 (HER2). According to the World Health Organization, breast cancer affected 2.3 million women worldwide in 2022, causing around 670,000 deaths, making it the second most common cancer globally [1][2]. In Asia, particularly in Malaysia, the incidence continues to rise, with breast cancer being the most frequently diagnosed cancer, accounting for 16.2% of all cases [3].

Classifying breast cancer into molecular subtypes is crucial for determining appropriate treatment strategies. Based on gene expression profiling, the main subtypes include Luminal A, Luminal B, HER2-enriched, Basal-like, and Claudin-low [4][5][6][7]. Each subtype exhibits distinct biological characteristics and responds differently to therapeutic interventions, making accurate classification critical for personalized care [8]. For instance, luminal tumors are typically treated with endocrine therapy, HER2-enriched tumors respond well to targeted antibody therapy, while Basal-like or triple-negative tumors, although aggressive, may show responsiveness to chemotherapy.

Currently, molecular subtyping is mainly done through immunohistochemistry (IHC), which requires invasive tissue biopsies. However, this approach presents several limitations, including patient discomfort, procedural risks, and diagnostic inaccuracies. Studies have reported that up to 30% of biopsies may yield false-negative results due to sampling errors, potentially delaying appropriate treatment. Although imaging modalities such as mammography, magnetic resonance imaging (MRI), and ultrasound provide valuable tumor-related information, they are not routinely utilized for molecular subtype classification. Among these, digital mammography stands out as a cost-effective and widely accessible screening tool, particularly in resource-limited settings. It also offers practical advantages in terms of acquisition time, typically requiring only 15 to 20 minutes compared to longer durations for ultrasound and MRI. Among these, digital mammography is particularly notable for its cost-effective and widespread accessibility, especially in resource-limited settings. It also offers practical advantages over other imaging techniques, with screening procedures typically taking just 15 to 20 minutes, compared to over 30 minutes for ultrasound and between 30 and 90 minutes for MRI scans [9]. Furthermore, mammography enables early detection of breast cancer by identifying abnormalities two to three years before they become clinically palpable [10].

Despite such advances, a substantial research problem remains, namely the lack of accurate, dependable, and non-

invasive techniques for the categorization of breast cancer genetic subtypes utilizing imaging data. Current diagnostic methods predominantly depend on invasive techniques, whereas contemporary imaging modalities primarily emphasize tumor identification rather than subtype classification. Furthermore, the incorporation of sophisticated artificial intelligence (AI) methodologies, especially deep learning architectures like Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), has not been thoroughly examined regarding non-invasive subtype classification. This deficiency constrains the advancement of effective computer-aided diagnostic systems that can assist in clinical decision-making. Non-invasive AI-based approaches have the potential to reduce unnecessary biopsies and improve diagnostic efficiency.

Therefore, the primary objective of this study is to systematically review and analyze the application of artificial intelligence techniques in breast cancer molecular subtype classification using imaging data. Specifically, this study aims to: 1) categorize existing approaches into CNN-based, Vision Transformer-based, and hybrid CNN–ViT models; 2) evaluate their performance, strengths, and limitations; and 3) identify current research gaps and future directions for improving non-invasive diagnostic methods.

The significance of this study lies in its contribution to advancing non-invasive breast cancer diagnosis by providing a comprehensive synthesis of state-of-the-art AI techniques. By critically analyzing existing methods, this review offers valuable insights into the effectiveness of different deep learning architectures and highlights opportunities for integrating imaging, radiomic features, and advanced models to enhance classification accuracy. Ultimately, this work supports the development of reliable, AI-driven diagnostic tools that can reduce dependence on invasive biopsies, improve early detection, and facilitate personalized treatment planning.

In this study, a systematic literature review based on the PRISMA 2020 framework is conducted to evaluate recent advancements in AI-driven breast cancer molecular subtype classification. The selected studies are analyzed and categorized according to their methodological approaches, including CNN-based models, Vision Transformer-based models, and hybrid architectures. Through this structured analysis, the study aims to provide a clear understanding of current trends, challenges, and future research opportunities in this domain.

This review bridges AI-based imaging analysis with clinically relevant molecular subtype classification, directly linking computational advancements to personalized treatment strategies and decision-making in breast cancer care.

II. Research Methodology

This systematic literature review adopts a focused (scoping) approach to explore AI's clinical applications in mammography imaging for classifying breast cancer molecular subtypes. The study specifically emphasizes recent and highly relevant works that address non-invasive subtype classification using imaging data, ensuring a targeted and in-depth analysis within a clearly defined research scope. Additionally, it examines other diagnostic methods, highlighting their current limitations and discussing future directions for AI in breast cancer diagnosis and treatment planning. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guiding principles were employed for the systematic review. This approach ensures a consistent and replicable process for identifying, selecting, and evaluating the relevant studies [11]. Fig. 1 presents the PRISMA flow diagram, illustrating the inclusion and exclusion process for articles considered in this literature review.

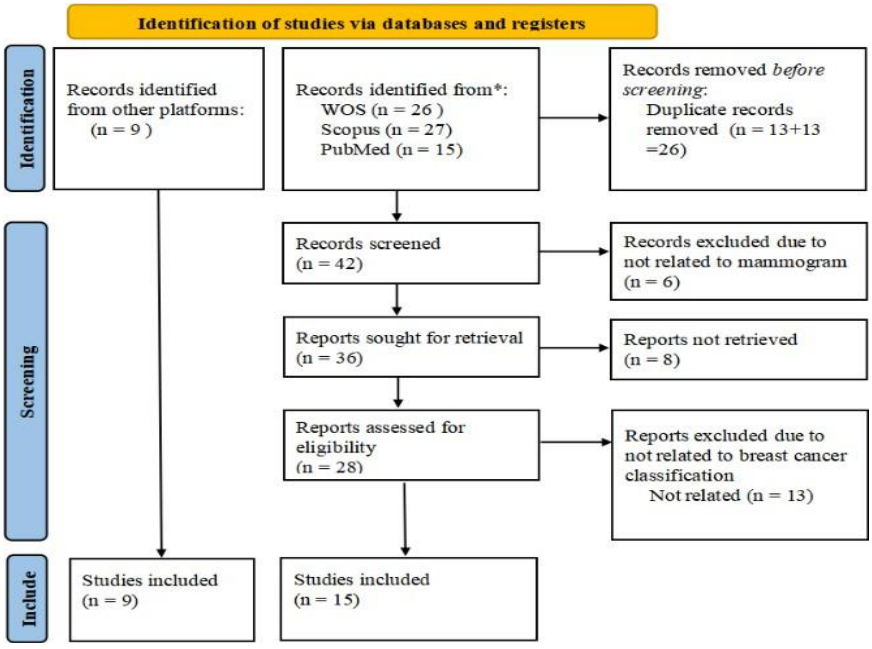


Fig. 1. PRISMA flow diagram.

A. Research Question

The research questions for this study are outlined in Table I, defining the overall purpose and expected outcomes. Each question is presented along with its motivation.

TABLE I. RESEARCH QUESTIONS AND MOTIVATIONS

Research Question	Main Motivation
What is the most effective image processing technique applied to extract tumor features from mammographic images?	To develop image processing technique to extract radiomic features, including shape and texture, relevant to breast cancer molecular subtypes from digital mammographic images.
What is the most effective model architecture for breast cancer molecular subtypes classification in terms of performance?	To construct classification models to analyze radiomic features and classify breast cancer into molecular subtypes.
How can a GUI be designed to support seamless interaction and decision-making for breast radiologists for identifying molecular subtypes of breast cancer?	To develop a Graphical User Interface (GUI) that allows users to automatically identify breast cancer molecular subtypes from digital mammographic images.

B. Search Scheme

The primary digital libraries used for this search were Web of Science (WOS), Scopus, and PubMed. These databases were selected for their comprehensive collections of peer- reviewed studies and their capability to offer access to the latest advancements in biomedical engineering and artificial intelligence applications. Additionally, several breast cancer classification journals were identified on platforms outside these three databases, including Google Search, Google Scholar, ScienceDirect, and IEEE Xplore Library.

C. Systemic Review Process

The systematic review followed the PRISMA framework, consisting of three key phases, which are identification, screening, and analysis. The literature search was conducted from 2020 to November 6, 2024, using the search string: "breast" AND ("cancer" OR "tumour") AND "subtype*" AND "deep learning" AND "mammogra*".

During the identification stage, a total of 68 studies were obtained from Web of Science (WOS), Scopus, and PubMed. However, 26 studies were removed after filtering for duplicate papers. Besides, an additional 9 studies were identified from other platforms. In the screening stage, the 27 papers were further excluded, as 6 studies were unrelated to mammograms, 8 studies could not be retrieved, and 13 studies were not focused on breast cancer classification. Finally, a total of 24 research articles were selected for review after performing an identification from other platforms. Furthermore, some studies that explored other modalities for breast cancer classification, such as Magnetic Resonance Imaging (MRI), ultrasound, and histopathology techniques, were considered as references. While some of these studies were included in the final set of 24 research articles, others were excluded but were still referenced in several sections of the literature review. Fig. 2 displays the distribution of the final set of selected research articles.

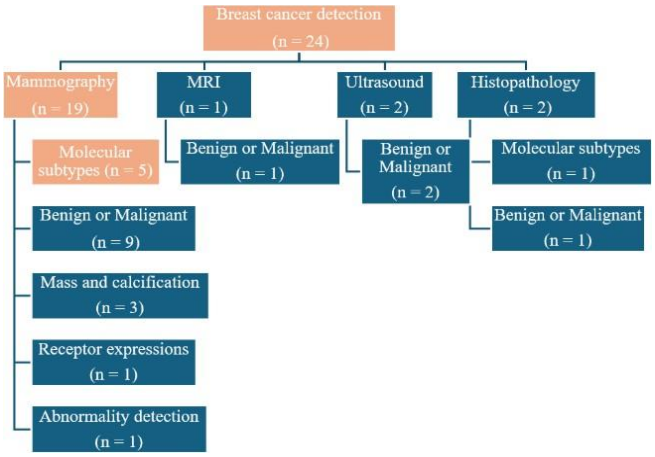


Fig. 2. Distribution of the final set of selected research articles.

D. Inclusion and Exclusion Principles

Table II lists the inclusion and exclusion principles for the chosen research. Stringent inclusion and exclusion criteria were applied to ensure the relevance and quality.

TABLE II. RESEARCH INCLUSION AND EXCLUSION CRITERIA

Inclusion criteria	Exclusion criteria
IC1 – Search will be conducted across all fields.	EC1 – Studies not related to breast cancer.
IC2 – Limited to open access and open archive materials.	EC2 – Studies lack focus on tumor radiomics feature extraction and classification.
IC3 – Articles and research articles.	EC3 – Review articles, books, chapters in book series, news articles, conference proceedings, case reports, and abstracts without full papers.
IC4 – The publication date range is set from 2020 to January 2026.	EC4 – Articles with inaccessible full texts.
IC5 – Studies published in English.	EC5 – Non-English publications.

III. RESULTS AND DISCUSSION

Table IV presents all the included studies, detailing their sample size, sample source, processing methods, architecture, and SWOT analysis (which examines strengths, weaknesses, opportunities, and threats).

Recent studies have explored the application of artificial intelligence techniques, particularly Convolutional Neural Networks (CNNs) and emerging Vision Transformer (ViT) architectures, for breast cancer classification using medical imaging. CNN-based approaches have demonstrated strong performance in tumor detection and classification tasks due to their ability to capture local spatial features. Meanwhile, Vision Transformer-based models have shown potential in modeling global contextual relationships within medical images. Hybrid CNN–Transformer architectures have also been introduced to leverage the strengths of both approaches.

However, most existing studies primarily focus on binary classification tasks, such as distinguishing between benign and malignant tumors, rather than performing detailed molecular subtype classification. Furthermore, many models rely on limited datasets or single imaging modalities, which may affect

generalizability and robustness. These limitations highlight the need for a comprehensive analysis of current AI-based methods specifically targeting non-invasive breast cancer molecular subtype classification. Therefore, this study aims to bridge this gap by systematically reviewing and comparing existing approaches while identifying key challenges and future research directions.

The overall workflow of AI-based breast cancer analysis adopted in this study is illustrated in Fig. 3, which presents a systematic framework encompassing data acquisition, preprocessing, feature extraction, model development, and classification.

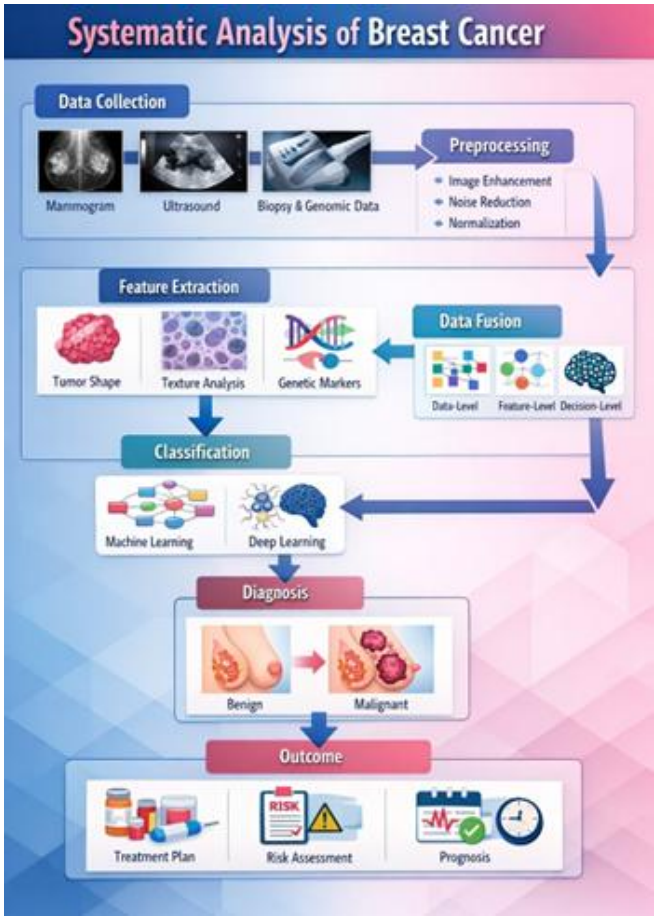


Fig. 3. Systematic framework for breast cancer analysis.

A. Machine Learning Approaches

Studies such as Khalid et al. [12] and Ma et al. [8] have employed machine learning approaches to classify mammogram images. Khalid et al. [12] implemented six classification models to distinguish between benign and malignant breast tumors, concluding that the Random Forest algorithm achieved the highest accuracy of 96.49%. On the other hand, Ma et al. [8] proposed a Naïve Bayes-based binary classification model to identify various molecular subtypes of breast cancer. These approaches, however, rely on manual feature selection and engineering, which necessitates human input and expertise. For example, the former utilized three modules, the low-variance features removal, univariate feature selection, and recursive feature elimination, to refine their models. Similarly, the latter combined manual segmentation with automated feature selection and extraction using LASSO (Least Absolute Shrinkage and Selection Operator) method.

B. Deep Learning Approaches

In contrast, deep learning approaches offer a significant advantage through their fully automated methodologies, from segmentation to classification. This is particularly beneficial in domains where the relevant features are not predefined or easily identifiable. For example, Carvalho et al. [13] introduced an end-to-end classifier leveraging SegNet and UNet for segmentation, thereby processing raw images without manual intervention. While deep learning typically requires large datasets for training, it is better suited for complex tasks such as image recognition, surpasses the traditional machine learning approaches.

C. Histopathology

Histopathology, which involves the diagnosis and study of diseases of the tissues level, plays a critical role in breast cancer analysis. It typically involves examining tissues or cells under a microscope and often requires a biopsy [14]. Biopsy is a procedure to remove cells, tissue or fluid from the targeted organ for examination. As the reference standard for breast cancer diagnosis, histopathology enables classification into molecular subtypes [15], which is essential for specialist to determine the type of breast cancer and devise an appropriate treatment plan.

Murtaza et al. [16] introduces a two-level hierarchical classifier employing various types of deep neural networks to classify breast cancer using histopathology images. The framework identifies eight distinct output classes, which are classifier distinguishes images as benign and malignant at the first level, while the second level further categorizes subtypes within each category. Among the architectures evaluated, the k-Nearest Neighbors (kNN) algorithm demonstrated the best performance, achieving 95.48% accuracy for first-level classification, 94.62% for benign subtypes, and 92.45% for malignant subtypes. Similarly, Rafiq et al. [17] utilizes histopathology images from the Kaggle dataset to evaluate the classification performance of a fusion-based convolutional neural network (CNN) framework. The authors proposed three CNN-based architectures for benign and malignant cancer classification, which are a straightforward CNN model (1-CNN), a fusion CNN model (2-CNN) and three CNN model (3-CNN). Among these, the 3-CNN architecture recorded the highest accuracy of 97.90%.

D. MRI and Ultrasound Imaging

Although histopathology remains a critical step in diagnosing breast cancer, non-invasive screening methods such as mammography, MRI and ultrasound have become increasingly favored as feasible options for large-scale screening. These methods offer advantages in terms of accessibility and reduced patient discomfort, often outweighing the invasive nature of histopathology in routine screening contexts. For instance, Carvalho et al. [13] utilized 20 dynamic contrast-enhancing magnetic resonance imaging (DCE-MRI) images from the public Quantitative Imaging

Network (QIN) database, recorded from 10 patients, to evaluate the accuracy of classifying tumor malignancy using the three-time-point (3TP) method, and the model achieved an impressive 100% classification accuracy. However, the small dataset used in the study raises concerns about potential overfitting, which could limit the model's generalizability. Beyond MRI, ultrasound imaging has also demonstrated utility in breast cancer classification. For example, Alshowarah [18] analyzed 780 ultrasound images from the Breast Ultrasound Images (BUSI) dataset to compare the performance of various classification architectures. Among these, the Random Forest algorithm outperformed others, achieving an accuracy of 88.63%. Similarly, T. Zhang et al. [19] and Jiménez-Gaona et al. [20] employed ResNet-50 and ResNet-18 architectures, respectively, to classify benign and malignant breast cancer using ultrasound and mammogram images, demonstrating the growing potential of advanced deep learning techniques in non-invasive diagnostics.

E. Mammography Imaging

Digital mammography is a cost-effective and widely available diagnostic tool, particularly in less-developed countries. It offers significant advantages over other imaging methods, as screening mammograms typically take only 15 to 20 minutes, compared to the 30 minutes or more required for ultrasound and the 30 to 90 minutes needed for MRI scans [9]. Additionally, mammography is particularly suited for early breast cancer detection, capable of identifying lumps two or three years before they can be felt by the patient [10]. Therefore, regular breast cancer screenings using mammography play a vital role in reducing modality risks and facilitating earlier and more effective treatment. Numerous studies have leveraged mammogram images for breast cancer classification. For example, Bobowicz et al. [21] used mammogram images from Medical University of Gdansk (MUG) and Chinese Mammography Database (CMMD) to train and test the models for classifying benign and malignant breast cancers. In their study, ResNet-34 demonstrated the best performance, achieving an accuracy of 82.2%. Mammography has also been employed in other applications, such as the classification of receptor expressions [22], feature extraction and early-stage breast abnormality detection [23], and the classification of breast masses and calcification [24]. These advancements highlight the versatility and effectiveness of mammography in breast cancer diagnosis and research.

F. Breast Cancer Molecular Subtypes Classification

As previously highlighted, breast cancer subtyping has critical therapeutic implications for clinical management of the disease, thus several studies have focused on developing classifiers for breast cancer molecular subtypes, including luminal A, luminal B, human epidermal growth factor receptor 2 (HER2)-enriched, and triple-negative types.

G. Convolutional Neural Network and Machine Learning

Molecular subtyping remains challenging due to subtle radiographic differences between subtypes, heterogeneity in tumor morphology, and frequent class imbalance in datasets (e.g., Luminal A dominance). Binary classifiers often fail to capture subtype-specific biomarkers, leading to accuracy <80% (Table IV). For instance, binary classification model which used to classify luminal A and non-luminal A subtypes is proposed by Liu et al. [25] with ResNet-50, achieving an accuracy of 53.3%. Similarly, Ma et al. [8] employed a Naive Bayes-based binary classification model to differentiate molecular subtypes recording accuracies of 0.796 for triple-negative, 0.748 for HER2-enriched and 0.788 for luminal subtypes. In contrast to these binary approaches, studies such as Mota et al. [26], Nissar et al. [15] and T. Zhang et al. [19] developed models capable of distinguishing multiple molecular subtypes. For example, Mota et al. [26] achieved an area under the curve (AUC) of 60.62% using a ResNet-101 deep CNN, Nissar et al. [15] reported an accuracy of 90% using MobileNet-V3 —highlighting that high accuracy may mask poor sensitivity in minority classes (e.g., HER2-enriched); and T. Zhang et al. [19] achieved an accuracy of 88.5% with ResNet-50. These multi-class classification model demonstrate the potential for more comprehensive diagnostic tools, advancing precision medicine in breast cancer treatment.

H. Vision Transformers

Despite the success of convolutional neural networks (CNNs) in medical imaging, their limited local receptive field restricts their ability to capture global context information [27]. A recent innovation, Vision Transformers (ViTs) has demonstrated a great potential to be an alternative to CNNs in image analysis due to their capacity to process large datasets and learn complex patterns. Gheflati & Rivaz [27] highlights the application of Vision Trans- formers (ViTs) in classifying breast ultrasound images, showcasing their ability to capture global dependencies within images through self-attention mechanisms by achieving 86.7% of accuracy, outperforming several state-of-the-art CNN architectures, including ResNet-50, VGG16, InceptionV3 and NASNetLarge. This study underscores ViTs' adaptability to small datasets using transfer learning, al- though computational requirements remain a challenge. Zeng et al. [28] also applied a Vision Transformer Multi-head Attention (ViT-MHA) architecture for breast mass diagnosis, leveraging bilateral mammography images to enhance diagnostic accuracy to 92.7%.

I. Hybrid CNN-ViTs

The hybridization of ViTs with CNNs has further improved image classification tasks. For example, Boudouh & Bouakkaz [29] combines ViTs for contextual feature extraction with CNNs for local detail recognition using the CBIS-DDSM dataset. ViTs outperform CNNs in capturing global tumor context but require 3–5× more training data and GPU resources. Hybrid ViT-CNN models (Boudouh & Bouakkaz, 2024 [29]) mitigate this at the cost of interpretability. Among the proposed models, the fusion of ViT++ and VGG16 achieved the highest accuracy of 99.22% underscoring the effectiveness of this hybrid approach. These studies highlight the potential of Vision Transformers in medical imaging while challenges such as computational demands and data diversity remain areas for future research and optimization.

J. Graphical User Interface

A Graphical User Interface (GUI) is a critical feature for transitioning a proposed classifier into clinical use. Zhang et al. [30] developed a web server using the open-source Dash framework to host their trained BSNet model. This web application enables data scientists to analyze mammogram images by uploading them to the server, which outputs probabilities for three breast disease classifications along with extracted features. Besides, Jiménez-Gaona et al. [20] created an open-source mobile app named "BraNet" for the segmentation and classification of benign and malignant breast cancer using mammography and ultrasound breast images. This app employs ResNet-18 deep learning algorithms, offering a practical tool for healthcare applications. Despite promising accuracy, GUI tools face regulatory (e.g., FDA/CE certification), interoperability (DICOM/HL7 integration), and physician-trust barriers.

K. Recommendations and Limitations

The literature review reveals that only a limited number of studies have specifically focused on developing classification models for breast cancer molecular subtypes using mammographic images. In particular, only five studies have directly addressed this problem, as summarized in Table III. This highlights a significant research gap in non-invasive subtype classification using imaging data.

To provide a broader perspective and enable cross-study comparison, a comprehensive summary of relevant studies, including model architectures, datasets, and performance metrics, is presented in Table IV.

Furthermore, to facilitate systematic comparison and evidence synthesis across studies, a standardized summary of included works is presented in Table V, highlighting model architectures, datasets, sample sizes, imaging modalities, and performance metrics.

TABLE III. SUMMARY OF CNN TECHNIQUES USING MAMMOGRAPHIC IMAGES TO CLASSIFY MOLECULAR SUBTYPES IN PAST STUDIES.

Author	Method	Accuracy	Remark
Mota et al. (2024) [26]	ResNet-101	60.62%	Low accuracy
Nissar et al. (2024) [15]	MobileNet-V3 with Convolutional Block Attention Module	90%	High complexity
T. Zhang et al. (2023) [19]	ResNet 50 with intra-modality and inter-modality	88.5%	High complexity
Ma et al. (2019) [8]	Naïve Bayes	75% to 80%	Low accuracy
Liu et al. (2023) [25]	ReNet 50	53.3%	Low accuracy

TABLE IV. SUMMARY OF THE PAST RESEARCH

Author	No. of sample and sample source	Processing methods	Architecture	Result	SWOT analysis (S = Strengths, W = Weaknesses, O = Opportunities, T = Treats)	Remark
Mammogram						
Mota et al., 2024 [26]	> 660 patients, 1397 images > Mammography images from OPTIMAM imaging database	> Cropping > Sampling for imbalanced data (imbalanced data, oversampling and under sampling) > Data augmentation > 3-fold cross-validation	ResNet-101 deep CNN	Binary classification: Best performance = Oversampling and data augmentation Average accuracy = 79.02% Average AUC = 64.69% multi-class classification: Best performance = Oversampling and data augmentation AUC = 60.62%	S = Access to a larger dataset of mammography images. Implementation of robust pre-processing and data augmentation techniques, enhancing model generalizability. W = Dependence on labelled mammographic data. Lack of data variability, potentially affecting the model's performance and generalizability. O = Potential to address class imbalance issues for more balanced model training. T = Exclusion of Ki-67 data may reduce the model's ability to differentiate between subtypes. Challenges in distinguishing subtle differences among molecular subtypes.	Classification molecular subtype
Nissar et al., 2024 [15]	> 2358 images before augmentation, 4987 images after augmentation > Mammography	> Image cropping and filtering > Image enhancement and augmentation > Image normalization	MobileNet-V3 architecture with a Convolutional Block Attention Module (MOB-CBAM)	Best performance = Median filter Accuracy = 0.90 Precision = 0.91 Recall = 0.90 F1-score = 0.89 Matthew's correlation	S = High accuracy. Capable of identifying various types of breast cancer, including masses, calcifications and benign or malignant tumour, as well as predicting molecular subtypes. W = Small sample size for certain classification tasks. High complexity of deep learning models.	Classification of benign or malignant tumour, as well as molecular subtypes

				coefficient = 0.85	O= Potential to integrate model-agnostic explainability methods, which would improve interpretability and assist in error and bias detection. T = Risks of errors or biases in predictions.	
T. Zhang et al., 2023 [19]	> 3360 paired cases > Mammography and ultrasound	> Feature extraction > Feature map > Intra-Modality Attention > Refined feature > Inter-Modality Attention	ResNet 50 with intra- and inter-modality attention modules	Accuracy = 88.5% MCC = 0.837	S = Combines multiple image views to improve classification of molecular subtypes. Employs multi-modal image analysis by integrating mammography and ultrasound. Uses intra- and inter-modality attention modules to effectively integrate features from different modalities, further boosting accuracy. W = Retrospective study design may lead to missing values. Small sample size within certain subgroups. Cases are sourced from a single centre, which may restrict the study's applicability across diverse populations. O = Potential to incorporate MRI into the multimodal models, which could further enhance the performance. T = Risks of model bias and errors.	Classification of molecular subtypes
Ma et al., 2019 [8]	> 331 patients, 662 images > Mammography	> Radiomic feature extraction > Normalization > LASSO feature selection > Oversampling technique > 10-fold cross validation	Naïve Bayes machine learning	Best performance = CC and MLO Accuracy: Triple-negative vs non-triple-negative = 0.796 HER2-enriched vs non-HER2-enriched = 0.748 Luminal vs non-luminal = 0.788	S = Combines multiple mammogram views (MLO and CC) to improve accuracy. Utilizes Naive Bayes and LASSO feature selection to enhance the prediction accuracy. W = Focuses on binary classification instead of multi-class classification. Limited dataset diversity may reduce the model's generalizability. Does not classify between luminal A and luminal B subtypes. Manual lesion segmentation. O = Develops multi-class classifier for molecular subtypes prediction. Incorporates larger and more diverse datasets. T = Less reliability due to data imbalance.	Classification molecular subtype
Liu et al., 2023 [25]	> 158 patients, 170 lesions > Mammography, MRI	> ROI labelling > ROI segmentation > Normalization	ResNet 50	Mammography: Accuracy = 53.3% Sensitivity = 66.7% Specificity = 44.4% AUC value = 0.593 (95%CI, 0.436-0.737) T2WI models = optimal	S = Multi-modal imaging approaches provide more comprehensive insights. W = Reliance on manual labelling of ROI. Only focus to Luminal A and non-Luminal A breast cancer. O = Opportunity to leverage ResNet model to classify all molecular subtypes. T = Small sample size could lead to model inaccuracies.	Classification Luminal A and non-Luminal A
Saber et al., 2021 [31]	> 322 images from MIAS > Mammography	> Noise removal > Histogram equalization > Morphological analysis > Threshold-based segmentation > Image resizing	> Inception V3 > ResNet 50 > VGG 19 > VGG 16 > Inception-V2 ResNet (Transfer learning)	Best performance = VGG 16 Accuracy = 98.96% Sensitivity = 97.83% Specificity = 99.13% Precision = 97.35%	S = Achieves high accuracy. Reduces training time and improves classification performance by automatic extracting of affected patch using segmentation. Compares many pre-trained CNN. W = The method focuses solely on benign and malignant cases, without addressing molecular subtypes critical for treatment decisions. Limited dataset diversity. O =	Classification between normal, benign and malignant

		> Data splitting > Data augmentation		F-score = 97.66% AUC = 0.995	Extends the function to classify molecular subtypes. T = Potential biases in classification outcomes as dataset lacks diversity.	
Khalid et al., 2023 [12]	> 1501 individuals, 3002 images (CC and MLO) > Mammography, kaggle.com	> Normalization and standardization > Data balancing > Feature extraction > Creating variants of photos > Noise reduction	Machine learning > Random Forest (RF) > Decision tree (DT) > k-nearest neighbors (KNN) > Logistic regression (LR) > Support vector classifier (SVC) > Linear support vector classifier (linear SVC)	Accuracy: RF = 96.49% DT = 93.86% LR = 92.98% KNN = 92.11% Linear SVC = 89.47% SVC 87.72%	S = High accuracy achieved for the majority of models. W = Classification limited to malignant and benign cases, rather than including molecular subtypes. Manual feature selection and engineering. O = Exploration of the RF model to classify molecular subtypes. T = Risk of false positives post-processing. Loss ability to discover complex patterns due to the manual engineered features.	Classification between benign and malignant
Bobowicz et al., 2023 [21]	> 789 patients, 1968 images from MUG > 1775 patients, 3744 images from CMMD > Mammography	> Data augmentation > Feature extraction > Patch features > 5-fold cross validation	> ImageNet (ResNet-18, ResNet-34, ResNet-50 and EfficientNet-B0) (Feature extraction) >Multiple instance learning-based	Best performance = ResNet-34 AUC-ROC = 0.892±0.023(-) F1-Score = 82.2±2.6(%) Accuracy = 82.2±2.6(%) Precision = 82.4±2.4(%) Recall = 82.2±2.6(%)	S = Access to two large datasets, which enhances accuracy and robustness. Integration of both CC and MLO views, leading to improved performance. Utilization of weakly supervised learning methods reduces reliance on fully annotated datasets. W = Computational complexity may increase deployment costs. O = Potential to expand classification capabilities to include molecular subtypes. T = The presence of pectoral muscle regions within image patches could introduce high-intensity pixel disturbances. Model bias due to data imbalances.	Classification between benign and malignant
M. Zhang et al., 2023) [30]	> 2321 images > Mammography cases	> Feature extraction > Data augmentation > 5-fold cross validation > Balanced sampling > Cross-entropy	Weakly supervised deep learning framework BSNet	Average AUC: Test set = 0.89 External validation set = 0.92	S = Compare BSNet model with others to prove its best performance. Weak supervision reduces dependency on extensive manual pixel-level annotations. Accessible via a web server. W = Limited comparison scope, as the model's fine-tuning performance is evaluated only on pre-trained models based on the ImageNet classification task. Detection capability is limited to breast cancer with HR status and benign tumours. O = Potential to expand classification capabilities to include molecular subtypes. Enhance model training and validation by collecting more extensive datasets. T = Robustness may be low. Lack of reliable benchmark for comparative analysis.	Detect HR status and benign
Mishra et al., 2023) [32]	> No indicated the number of images > Digital database for screening mammography (DDMS)	> Cleaning the dataset > Data exploratory analysis > Data distribution analysis	> Logistic regression > Neural network architecture with 4 hidden layers	Best performance = Neural network architecture with 4 hidden layers Train accuracy	S = Achieves high accuracy in detecting breast cancer. W = The model does not classify molecular subtypes, which are crucial for identification of treatment. Unbalanced dataset.	Classification between benign and malignant

	> Mammography	> Data encoding > Dataset splitting > Feature scaling		= 98.99% Test accuracy = 98.83%	O = Potential to extend the proposed method for classifying molecular subtypes. Incorporating more diverse and reliable datasets to improve sensitivity and overall robustness. T = Risk of bias due to dataset imbalances.	
Harish et al., 2024 [33]	> 30 different features, 570 data > CT mammogram from Kaggle	> Data preprocessing (resizing images and normalizing pixel values) > Feature localization and extraction	Convolutional neural network (CNN) equipped with a region proposal network, RPN	Accuracy = 89%	S = A faster, sequential CNN model is proposed for breast cancer classification. W = Limited dataset used, with only 50 malignant and 50 benign cases for training. Focuses only on classification of normal, malignant and benign cases. O = Potential to enhance the accuracy, precision and recall by integrating with other deep learning model without increasing time complexity. Opportunity to expand the method to classify molecular subtypes of breast cancer. T = The small dataset increases the risk of overfitting, which may affect the performance on unseen data. Potential for reduced accuracy in real-world applications.	Classification of normal, malignant and benign
Agnes et al., 2020 [34]	> 322 images from MIAS database > 4500 images after data augmentation > Mammography	> Noise and background removal > Data normalization > Data augmentation > Feature reduction	Multiscale All Convolutional Neural Network (MA-CNN)	> Accuracy = 96.47% > Sensitivity = 96% > AUC = 0.99	S = Highest in accuracy compared with others. A larger stride convolution operation is used for dimension reduction, which preserves the neighborhood pixel relationship and eliminates the need for separate pooling layer. Implements multiple dilated convolution operation to extract multiscale features for classification. Operates as an end-to-end system, eliminating the manual feature engineering from the raw images. W = Limited dataset diversity as it relies on a single dataset only, which may affect the model's generalizability. O = Incorporating patch details to enhance the accuracy of the current classifier model. Expands the model to classify molecular subtypes. Training on larger and more diverse datasets to enhance reliability. T = May result in overfitting as reliance on a single dataset.	Classification between normal, benign and malignant
Jiménez-Gaona et al., 2024 [20]	> 5892 images > BUSI, collected by Rodrigues et al., UDIAT diagnostic centre, CBIS-DDSM, mini-MIAS, Inbreast, VinDr-Mammo > Mammography and ultrasound images	> Data normalization and automatic ROI annotation > Data augmentation > ROI segmentation	> SNGAN (generate synthetic images) > ResNet-18 (image classification) > SAM model (image segmentation)	Accuracy: US = 94.7% benign and 93.6% malignant DM (I) = 80.9% benign and 76.9% malignant DM (II) = 73.7% benign and 72.3% malignant	S = Large and diverse datasets. Offers an open-source API for mobile breast image processing and classification. W = Limited output provided, only benign and malignant classification is available. Less accurate for mammography images. Impacted by data quality. O = Integration with advanced models to improve accuracy and reliability. Expands the function to classify molecular subtypes. T = Risks of data privacy. Ethical challenges in AI.	Classification between benign and malignant

Boudouh & Bouakkaz, 2024 [29]	> 1871 images from CBIS-DDSM > Mammography	> Balance collected data > Noise reduction and enhancement filters > Data augmentation > Feature extraction and classification	> Vision Transformer, ViT++ (Contextual features) > CNN based on transfer learning techniques includes Xception, VGG16, RegNetX002 (Visual features)	Best performance = ViT++ with VGG16 Accuracy = 99.22%	S = Combines ViT++ for capturing global context with CNNs for extracting local spatial details. High accuracy. Employs pre-processing techniques to reduce noise and enhance the quality of input data. Offers flexibility in model design and provides a robust feature representation. W = High computational requirements for the complex fusion approach. The feature extraction phase does not include a feature selection step. Limited datasets diversity. O = Incorporates other pre-training models in the CNN branch. Expands to diverse datasets. Extends the fusion approach proposed to classify molecular subtypes. T = High costs of implementation. Risk of redundancy and increased computational load due to irrelevant or overlapping features.	Classification breast calcifications to distinguish between benign and malignant cases
Boudouh & Bouakkaz, 2023 [24]	> No indicated the number of images > CBIS-SSSM collection > Mammography	> Noise removal > Image enhancement > Data augmentation > Feature extraction (transfer learning techniques)	> ResNet 50 V2 > Xception (Feature extractor)	Best performance = Xception Accuracy = 99.99% Recall = 100% Specificity = 100% AUC = 1.00	S = A high accuracy method proposed. Employs advanced pre-processing techniques such as Non-Local Means, Gaussian and Median filters to remove noise. W = Unbalanced dataset, which is lack of data on the Calc class. The model does not classify molecular subtypes, which are crucial for identification of treatment. O = Potential to improve accuracy by adding another dataset or hybridization of pre-trained models for feature extraction. Expand the proposed method to classify the molecular subtypes. T = The model may struggle to detect breast Calc.	Classification between breast Mass and breast Calc
Huang et al., 2022 [35]	> 322 images from MIAS > Mammography	> Image normalization and denoising > Image enhancement (sharpening) and histogram modification > Segmentation of ROI > Feature extraction	SVM (Support vector machine)	Best performance = Gaussian radial basis kernel function (RBF) Average accuracy = 78.89%	S = Employs kernel functions for optimal classification performance. The method extracts diverse features including grayscale, geometric and texture properties. W = Dependency on labelling information. The training time increases with larger datasets. Requires extensive experimentation to determine kernel function parameters. Limited dataset diversity as it relies on MIAS dataset only. O = Combines with other effective techniques can enhance classification performance. Includes larger and more diverse datasets to improve model's reliability. Opportunity to expand the method to classify molecular subtypes of breast cancer. T = Risk of biased results if critical features are mislabeled or omitted. Long training time may require.	Detection of breast calcification clusters and classification of masses
Zeng et al., 2024 [28]	> 1003 mass and 694 normal cases from DDSM	> Input embedding > multi-head attention > Addition and	ViT's MHA	Accuracy = 0.927	S = Integrates bilateral information from both the left and right breast images to improve the diagnostic accuracy of breast mass detection. Eliminates the need of manual intervention by	Classification between mass and normal tissue

	> Mammography	normalization > Feed-forward			automating the process of feature extraction and classification. Overcomes registration difficulties by adopted ViT's MHA architecture. W = Limited datasets diversity. Fails to fully exploit the symmetric region details of the images. High complexity in bilateral image integration. O = Expands to diverse datasets. Extends the method proposed to classify molecular subtypes. T = Low performance on unilateral images.	
Ueda et al., 2021 [22]	> 1448 images (data set) > 225 images (test data) > Mammography	> Ground truth labelling > Image cropping and resizing > Binary cross-entropy > Data augmentation > Five-fold cross-validation	> VGG 16 > Inception V3 > ResNet 52 >DenseNet 121 (Combination)	Best performance = Inception V3 + ResNet 52 and VGG 16 + Inception V3 AUC: ER+ or - = 0.67 (0.58-0.76) PgR+ or - =0.61 (0.53-0.68) HER2-enriched or -not enriched = 0.75 (0.68-0.82)	S = Trained model is available online with Apache License 2.0. Blending multiple models is used to improve the predictive performance. W = Insufficient number of mammograms. Used the results from the preoperative biopsy report rather than the postoperative tissue report for patients with NAC as NAC can alter receptor expressions. O = Adopted visualization technology such as gradient-weighted class activation mapping. T = Lower accuracy.	Classification of receptor expressions
Thangavel et al., 2024 [23]	> No indicated the number of images > Mammography	> Adaptive fuzzy based median filtering > Noise removal > Normalization > Image enhancement > Feature extraction	> ResNet (Hierarchical feature learning) > U-Net (Segmentation)	Accuracy = 99% Precision = 98.6% Recall = 99.01% Specificity = 98.9%	S = A high accuracy feature extraction and early-stage identification model by combining two deep learning models. W = The model not involve in the classification of molecular subtypes, which is important for identification of treatment. O = Potential to use the proposed method to classify the molecular subtypes. Integrating the model with real-time data streams. T = May face challenges related to interpretability, data dependency, computational complexity and ethical concerns regarding patient privacy and consent when it integrated into clinical practice.	Feature extraction and early-stage breast abnormality detection
Magnetic Resonance Imaging (MRI)						
Carvalho et al., 2024 [13]	> 20 DCE-MRI, 10 patients from QIN > MRI image	> Image segmentation > Image registration	> SegNet and Unet (segmentation) > Three-time-point, 3TP (classification)	Best performance = Unet (0.9799 IoU, 0.9332 Dice) Accuracy for classification = 100%	S = Fully automated tumor segmentation and classification methodology. Achieves high accuracy. Use three-time-point (3TP) approach to enhance dynamic contrast behavior. W = Small dataset may reduce generalizability. High number of parameters required for deep learning models increases complexity. Long training time. O = Incorporating larger datasets to enhance reliability. Use optimization methods to obtain the best parameters for the architecture. Expands to include other imaging modalities such as mammography. Expand classification to include molecular subtypes could enhance treatment planning.	Classification the malignancy of the tumors into malignant type III (WashOut) and malignant type II (Plateau)

					T = Risks of overfitting. Less reliability.	
Ultrasound						
Alshowarah, 2023 [18]	> 780 images from Dataset of Breast Ultrasound Images (Dataset BUSI) > 5872 images after data augmentation >Ultrasound images	> Feature extraction > Data Augmentation, Statistical operations (Avg, Min, Max and fusion of fc6 and fc7)	> VGG-19 (features extraction) > KNN, Naïve Bayes, Random Forest, Decision Tree (classifiers)	Best performance = Random Forest Accuracy = 88.63% Precision = 0.88 Recall = 0.88 F-Measure = 0.88	S = Comprehensive comparison of multiple classifiers. Exploration of multiple scenarios in the training dataset. Application of various statistical operations to generate more datasets for better learning of CNN. W = Limited dataset diversity as it relies on a single dataset only, which may affect the model's generalizability. The method focuses solely on benign and malignant cases, without addressing molecular subtypes critical for treatment decisions. Insufficient feature exploration. O = Potential to improve the classification accuracy and reduce training time by conducting more investigation and using different algorithms. Extend the proposed method to include molecular subtypes classification. Incorporating different imaging modalities such as mammography. T = Risk of bias due to class imbalances in the dataset. Limited testing in diverse clinical environments.	Classification between benign and malignant
Gheflati & Rivaz, 2022 [27]	> 780 images from Dataset of Breast Ultrasound Images (Dataset BUSI) > 163 images from dataset B >Ultrasound images	> 5-fold cross-validation > SoftMax activation function (CNN TL) > Fine-tuning, weighted cross-entropy loss (ViT) > Data augmentation	> Vision Transformer (ViT) > State-of-the-art CNN includes ResNet-50, VGG16, InceptionV3, NASNetLarge	Best performance = ViT-B/32 (BUSI+B) Accuracy = 86.7% AUC = 0.95	S = ViT architecture excels in capturing global dependencies in images through its self-attention mechanism. Comprehensive comparison of ViT and multiple CNNs. ViT adapted well to small datasets. Diverse datasets used. W = High computational requirements for ViT. O = Combines ViT with other imaging modalities such as mammography. Expands the method proposed to classify molecular subtypes. T = High costs for implementation.	Classification of benign, malignant or normal
Histopathology						
Murtaza et al., 2020 [16]	> 58 patients, 1339 images from BreakHis > Histopathology image	> Data collection > Image augmentation > Feature extraction > Feature reduction	> AlexNet (Feature extraction) > KNN, SVM, NB, DT, LDA, LR	Best performance = k-Nearest Neighbors (kNN) First-level classifier = 95.48% Second-level classifier = 94.62% (Benign subtypes) 92.45% (Malignant subtypes)	S = The BMIC_Net model proposed is less complex, computationally efficient, reliable and more accurate. Eliminates the need for handcrafted feature extraction. Employs a two level hierarchical classifier, with eight distinct output classes are identified where first level classifies images as benign and malignant, second level further classifies the subtypes of benign or malignant. Feature reduction techniques is used to improve accuracy. W = Limited dataset diversity may reduce the model's generalizability. Dataset imbalance between benign and malignant subtypes. Invasive method as histopathological images	Classification between benign and malignant, and their subtypes

					are required. O = Expand to include imaging modalities such as mammography or MRI. Expand classification to include molecular subtypes could enhance treatment planning. T = Risk of bias. Limited testing in diverse clinical environments.	
Rafiq et al., 2023 [17]	> 220,025 images from Kaggle dataset > Histopathology image	> Data augmentation > Image segmentation	> Straightforward CNN model (1-CNN) > Fusion CNN model (2-CNN) > Three CNN model (3-CNN)	Best performance = 3-CNN Accuracy = 97.90% Recall = 96.80% Precision = 96.90% f1-score = 97.20%	S = Utilizes hybrid CNNs to improve accuracy and optimize training efficiency. Utilizes a larger dataset. W = High computational complexity of fusion approach. Limited dataset diversity may reduce the model's generalizability. Relies on invasive biopsy method to obtain histopathological images. O = Expands dataset with external source. Extends the function to classify molecular subtypes. Incorporate other imaging modalities such as mammography. T = Risk of overfitting. Potential biases in classification outcomes as dataset lacks diversity.	Classification between benign and malignant

TABLE V. SUMMARY OF AI-BASED BREAST CANCER MOLECULAR SUBTYPE CLASSIFICATION STUDIES

Study	Year	Modality	Dataset Size	Model	Category	Task	Performance	Key Insight
Mota et al. [26]	2024	Mammography	1397 images	ResNet-101	CNN	Multi-class subtype	Acc: 79.02%, AUC: 64.69%	Data balancing improves performance
Nissar et al. [15]	2024	Mammography	4987 images	MobileNet V3 + CBAM	CNN	Subtype	Acc: 90%	Attention improves feature learning
T. Zhang et al. [19]	2023	Mammo + US	3360 cases	ResNet50 + Attention	CNN	Subtype	Acc: 88.5%	Multi-modal enhances accuracy
Ma et al. [8]	2019	Mammography	662 images	Naïve Bayes	ML	Binary subtype	Acc: ~79%	Limited to binary classification
Liu et al. [25]	2023	Mammo + MRI	170 lesions	ResNet50	CNN	Partial subtype	Acc: 53.3%	Small dataset limits performance
Ueda et al. [22]	2021	Mammography	1448 images	CNN ensemble	CNN	Receptor classification	AUC: up to 0.75	Limited subtype accuracy

The results presented in Table III indicate significant variability in classification performance across different models. CNN-based architectures such as MobileNet-V3 and ResNet-based models demonstrate relatively higher accuracy, particularly when enhanced with attention mechanisms or multi-modal inputs. However, these approaches often involve increased computational complexity. In contrast, traditional machine learning methods such as Naïve Bayes exhibit lower performance, highlighting their limitations in capturing complex feature representations required for multi-class subtype classification.

Furthermore, several studies report relatively low accuracy, particularly in cases with limited datasets or partial subtype classification, suggesting that data availability and model design remain critical challenges. These findings emphasize the need for more advanced architectures, such as hybrid CNN–Transformer models, which can effectively capture both local and global features to improve classification performance.

These existing models face several limitations. First, most primarily classify tumors as benign or malignant, which is insufficient for guiding precise treatment decisions tailored to breast cancer subtypes. Besides, some proposed models are overly complex and lack interpretability, making them less intuitive and harder to understand. This hinders their usability and acceptance in clinical settings. Further research is needed to develop a highly accurate yet less complex classifier for classifying breast cancer molecular subtypes using mammographic images.

Moreover, different studies employed different image processing methods. This inconsistency highlights the need

for further research to assess the effectiveness of each technique in improving segmentation and feature extraction process.

Despite these advancements, many studies face challenges related to bias, reliability and overfitting due to the small or imbalanced datasets. To address these issues, larger and more diverse datasets are essential for effectively training and testing models. In addition, data augmentation techniques, which generate new data samples from preexisting data can significantly improve model optimization and generalizability, making AI-based tools more robust and applicable in clinical settings.

IV. Conclusion

This review highlights the growing impact of artificial intelligence, in breast cancer classification using mammographic imaging. While AI-based models have shown promise in improving diagnostic accuracy and distinguishing between molecular subtypes, existing approaches still face challenges related to model complexity, interpretability, and dataset diversity. A majority of studies primarily focus on benign versus malignant classification, limiting their clinical utility in precision treatment planning. In addition, heterogeneity in preprocessing pipelines techniques underscore the need for standardized methodologies in order to enhance segmentation and feature extraction. To fully leverage AI in breast cancer diagnostics, future research should focus on developing more interpretable, highly accurate classification models while ensuring their seamless integration into clinical workflows. The adoption of larger, more diverse datasets and advanced machine learning techniques, such as Vision Transformers and hybrid CNN approaches, may potentially result in further improved classification performance. Ultimately, AI-driven solutions hold great potential in revolutionizing breast cancer detection, subtyping, and personalized treatment strategies.

Acknowledgment

The author acknowledges the support of the Ministry of Higher Education (MoHE) Malaysia through the Fundamental Research Grant Scheme (FRGS) under grant number FRGS-EC/1/2024/ICT02/UNIMAP/02/8.

References

[1] World Health Organization, “Breast cancer.” Accessed: Oct. 22, 2024. [Online]. Avail- able: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] World Cancer Research Fund, “Breast cancer statistics.” Accessed: Jan. 09, 2025. [On- line]. Available: https://www.wcrf.org/cancer-trends/breast-cancer-statistics/

[3] Global Cancer Observatory, “Cancer Today.” Accessed: Oct. 22, 2024. [Online]. Available: https://gco.iarc.fr/today/en

[4] C. M. Perou et al., “Molecular portraits of human breast tumours,” Nature, vol. 406, no. 6797, pp. 747–752, Aug. 2000, doi: 10.1038/35021093.

[5] “Comprehensive molecular portraits of human breast tumours,” Nature, vol. 490, no. 7418, pp. 61–70, Oct. 2012, doi: 10.1038/nature11412.

[6] J. I. Herschkowitz et al., “Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumours,” Genome Biol, vol. 8, no. 5, p. R76, May 2007, doi: 10.1186/gb-2007-8-5-r76.

[7] S. Łukasiewicz et al., “Breast Cancer—Epidemiology, Risk Factors, Classification, Prog- nostic Markers, and Current Treatment Strategies—An Updated Review,” Cancers (Basel), vol. 13, no. 17, p. 4287, Aug. 2021, doi: 10.3390/cancers13174287.

[8] W. Ma et al., “Breast Cancer Molecular Subtype Prediction by Mammographic Radiomic Features,” Acad Radiol, vol. 26, no. 2, pp. 196–201, Feb. 2019, doi: 10.1016/j.acra.2018.01.023.

[9] C. Okonkwo, “Breast MRI vs Mammogram vs Ultrasound: Which Imaging of the Breast Should You Choose?,” scan.com. Accessed: Nov. 20, 2025. [Online]. Available: https://uk.scan.com/health-hub/breast-mri-vs-mammogram-vs-ultrasound-which-imaging-of-the-breast-should-you-choose

[10] NSBSP, “Benefits and Limitations of Mammograms.” Accessed: Nov. 20, 2025. [Online]. Available: https://nsbreastscreening.ca/benefits-and-limitations-mammograms

[11] M. Nasser and U. K. Yusof, “Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction,” Diagnostics, vol. 13, no. 1, p. 161, Jan. 2023, doi: 10.3390/diagnostics13010161.

[12] A. Khalid et al., “Breast Cancer Detection and Prevention Using Machine Learning,” Diagnostics, vol. 13, no. 19, p. 3113, Oct. 2023, doi: 10.3390/diagnostics13193113.

[13] E. D. Carvalho, O. P. da Silva Neto, and A. O. de Carvalho Filho, “Deep learning-based tumor segmentation and classification in breast MRI with 3TP method,” Biomed Signal Process Control, vol. 93, p. 106199, Jul. 2024, doi: 10.1016/j.bspc.2024.106199.

[14] Cleveland Clinic, “Biopsy.” Accessed: Nov. 15, 2024. [Online]. Available: https://my.clevelandclinic.org/health/diagnostics/15458-biopsy-overview

[15] I. Nissar et al., “MOB-CBAM: A dual-channel attention-based deep learning generalizable model for breast cancer molecular subtypes prediction using mammograms,” Comput Methods Programs Biomed, vol. 248, p. 108121, May 2024, doi: 10.1016/j.cmpb.2024.108121.

[16] G. Murtaza, L. Shuib, G. Mujtaba, and G. Raza, “Breast Cancer Multi-classification through Deep Neural Network and Hierarchical Classification Approach,” Multimed Tools Appl, vol. 79, no. 21–22, pp. 15481–15511, Jun. 2020, doi: 10.1007/s11042-019-7525-4.

[17] A. Rafiq et al., “Detection and Classification of Histopathological Breast Images Using a Fusion of CNN Frameworks,” Diagnostics, vol. 13, no. 10, p. 1700, May 2023, doi: 10.3390/diagnostics13101700.

[18] S. A. Alshowarah, “Breast Cancer Detection System using Deep Learning Based on Fusion Features and Statistical Operations,” 2023. [Online]. Available: www.ijacsa.thesai.org

[19] T. Zhang et al., “Predicting breast cancer types on and beyond molecular level in a multi-modal fashion,” NPJ Breast Cancer, vol. 9, no. 1, p. 16, Mar. 2023, doi: 10.1038/s41523-023-00517-2.

[20] Y. Jiménez-Gaona et al., “BraNet: a mobile application for breast image classifica- tion based on deep learning algorithms,” Med Biol Eng Comput, vol. 62, no. 9, pp. 2737–2756, Sep. 2024, doi: 10.1007/s11517-024-03084-1.

[21] M. Bobowicz et al., “Attention-Based Deep Learning System for Classification of Breast Lesions—Multimodal, Weakly Supervised Approach,” Cancers (Basel), vol. 15, no. 10, p. 2704, May 2023, doi: 10.3390/cancers15102704.

[22] D. Ueda et al., “Training, Validation, and Test of Deep Learning Models for Classification of Receptor Expressions in Breast Cancers From Mammograms,” JCO Precis Oncol, no. 5, pp. 543–551, Nov. 2021, doi: 10.1200/PO.20.00176.

[23] Y. Thangavel, H. Garg, M. Alagarsamy, and D. Pradeep, “Revolutionizing breast cancer diagnosis . . . digital mammogram-based feature extraction,” Biomed Signal Process Control, vol. 94, p. 106268, Aug. 2024, doi: 10.1016/j.bspc.2024.106268.

[24] S. S. Boudouh and M. Bouakkaz, “Breast Abnormalities Classification Using Preprocessing and Deep Transfer Learning Techniques,” IET Conference Proc., pp. 53–58, 2023, doi: 10.1049/icp.2024.0900.

[25] M. Liu et al., “Identification of Luminal A breast cancer by using deep learn- ing analysis based on multi-modal images,” Front Oncol, vol. 13, Nov. 2023, doi: 10.3389/fonc.2023.1243126.

[26] A. M. Mota, J. Mendes, and N. Matela, "Breast Cancer Molecular Subtype Prediction: A Mammography-Based AI Approach," Biomedicines, vol. 12, no. 6, p. 1371, Jun. 2024, doi: 10.3390/biomedicines12061371.

[27] B. Gheflati and H. Rivaz, "Vision Transformers for Classification of Breast Ultra- sound Images," in 44th Annual Int. Conf. IEEE EMBC, 2022, pp. 480–483, doi: 10.1109/EMBC48229.2022.9871809.

[28] T. Zeng et al., "Vision Transformer-Based Breast Mass Diagnosis in Mam- mography Using Bilateral Information," in IICAIET, 2024, pp. 147–152, doi: 10.1109/IICAIET62352.2024.10730097.

[29] S. S. Boudouh and M. Bouakkaz, "Advancing precision in breast cancer detection: a fusion of vision transformers and CNNs," Applied Intelligence, vol. 54, no. 17–18, pp. 8170–8183, Sep. 2024, doi: 10.1007/s10489-024-05619-3.

[30] M. Zhang et al., "Developing a weakly supervised deep learning framework for breast cancer diagnosis with HR status based on mammography images," Comput Struct Biotechnol J, vol. 22, pp. 17–26, Jan. 2023, doi: 10.1016/j.csbj.2023.08.012.

[31] A. Saber et al., "A Novel Deep-Learning Model for Automatic Detection and Classifica- tion of Breast Cancer Using Transfer Learning," IEEE Access, vol. 9, pp. 71194–71209, 2021, doi: 10.1109/ACCESS.2021.3079204.

[32] P. Mishra et al., "Classification of Breast Cancer using a Novel Neural Network-based Architecture," in ICCCNT, 2023, pp. 1–6, doi: 10.1109/ICCCNT56998.2023.10306841.

[33] M. N. Harish et al., "CNN-Based Breast Cancer detection using Noninvasive methods," in ICBSII, 2024, pp. 1–5, doi: 10.1109/ICBSII61384.2024.10564089.

[34] S. A. Agnes et al., "Classification of Mammogram Images Using Multiscale All Con- volutional Neural Network (MA-CNN)," J Med Syst, vol. 44, no. 1, Jan. 2020, doi: 10.1007/s10916-019-1494-z.

[35] Y. Huang et al., "Application of Support Vector Machines for Breast Calcification Clus- ter Detection and Mass Classification," J. Phys. Conf. Ser., 2022, doi: 10.1088/1742- 6596/2400/1/012003.

Integrating Big Data and Machine Learning for Effective Cyberattack Prediction in e-Health Information Systems

Mohamed Abdelbaki[1]*, Latif Adnane[2], Charaf Eddine Ait Zaouiat[3]
Information Technology and Modelling Team-ENSA Marrakech, Cadi Ayyad University, Marrakech, Morocco[1, 2]
Polydisciplinary Faculty of Sidi Bennour, Chouaib Doukkali University, ELJADIDA, Morocco[3]

***Abstract*—This study proposes an intrusion-prediction framework for e-Health information systems that combines structured web-log analysis, supervised machine learning, and Apache Spark-based distributed processing. A corpus of 1,000,000 labeled HTTP log instances collected from a university hospital web environment was preprocessed into security-relevant features, including request method, request/response type, packet size, status code, URL length, and parameter count. Using a stratified 80/20 train-test split and five-fold cross-validation on the training data, we compared K-Nearest Neighbors (KNN), Logistic Regression, and Decision Trees. KNN achieved the best held-out performance, with 95.66% accuracy, 91.79% precision, 93.93% recall, 92.85% F1-score, and a 3.60% false positive rate. Logistic Regression and Decision Trees reached accuracies of 85.30% and 83.20%, respectively. Spark also reduced runtime substantially at the 1,000,000-instance scale, lowering KNN processing time from 12.0 s to 6.5 s. The results show that combining big data infrastructure with carefully tuned machine learning can improve both detection quality and operational feasibility in hospital cybersecurity monitoring.**

***Keywords*—Artificial intelligence; big data; cybersecurity; hospital information systems; log files**

I. Introduction

This exponential development of information and communication technologies rests on the increased utilization of more and more connected devices and tools in recent years. Through such evolution, users have easily and rapidly accessed information [1]. However, in the process, it has also facilitated illicit access to sensitive resources by malicious actors. The big challenge in the detection and prediction of attacks on communication systems, especially for hospital information systems, electronic medical records, and electronic health records, is related to the detection and prediction of such attacks [2]. Among the most attacked resources are those provided by web applications used in healthcare environments [3]. The log files record every event happening during system operation and are large in volume and complex in structure. Analysis of such log files is, therefore, of prime importance for addressing the risks involved with system attacks that may affect patient data and overall performance. Because of the huge size of these log files, it is not possible to conduct their manual analysis; extended processing times need to be used [4]. The novelty of this work lies in designing a cyberattack prediction framework that balances accuracy, scalability, and robustness to imperfect training data. Because security log datasets may contain noisy, inconsistent, or manipulated records, robustness becomes an important consideration alongside predictive performance [35]. This work will focus on the accuracy of different predictive algorithms, along with their time complexity, to provide an overall framework that will help administrators in healthcare identify and mitigate vulnerabilities effectively. The time complexity comparison analysis taken by algorithms used secures the development of an efficient solution to be able to handle volumes and improve efficiency, while personal data on patients is protected.

Also, by embedding Apache Spark as a big data processing engine, the model increases its speed of processing information and thus allows real-time analysis of large volumes of log data; in our evaluation, Spark supports distributed processing across scales up to 1,000,000 instances, addressing the practical scalability constraints of security analytics over large healthcare log volumes. This establishes a dual emphasis on predictive accuracy and computational efficiency, positioning the research as a practical contribution to improving the security posture of healthcare web applications.

The rest of this study is organized as follows. First, some related works concerning log file analysis and cybersecurity techniques are reviewed. Following this, the advantages and challenges in log file analysis applied to healthcare are discussed. Then, Big Data technologies and their importance in regard to security issues are described, and afterwards, machine learning tools and their practical applications in log analysis. The proposed approach and the implementation follow together with discussions of results.

II. Related Work

Basically, log file analysis refers to the intended examination and analysis of log files [5]. Log files are very important in tracking mistakes in data transmission, as well as tracking firewall activities [6]. Medical log file analysis is one of the most vital analyses that can be used in enhancing hospital information systems and electronic health records (EHR) [7]. Log files record not only the interaction of visitors with web applications but also trace various technical errors related to networks, software, and components, most of all, though, underlying security issues regarding patient data [8, 9]. However, the analysis of security logs is complex and requires advanced technical skills related to Big Data, Machine Learning, and both [30].

*Corresponding author.

Several works have described how Big Data and Machine Learning can be applied in security log file analysis. Landauer et al. [10] proposed a dynamic anomaly detection approach using clustering mechanisms to enable a self-learning algorithm to detect anomalies related to temporal behavior. On the other hand, Zhong et al. [11] proposed a deep learning log analysis-based intrusion detection approach based on the application of LSTM networks. Skopik et al. [12] investigated a unified pipeline that realizes several machine learning algorithms to analyze the system behavior and to identify deviations from established patterns, which in turn allows identifying both known and unknown attacks. Recent IDS surveys also show that Machine Learning and Deep Learning-based intrusion detection must address dataset quality, evasion techniques, false positives, and the detection of evolving attacks, which remains highly relevant for healthcare web environments [33, 37].

In the Big Data log file analysis context, Azizi et al. [13] proposed an approach where MapReduce could be used for the analysis of security-related log files concerning certain types of attacks, such as SQL injection or DDoS. Jeon et al. [14] proposed a new Big Data-based security logging system that extends security intelligence by analyzing data events created in enterprise log files, covering systems, application services, and IT infrastructure.

Despite these developments, most of the earlier studies have shown suboptimal performance because of data pre-processing deficiencies, which reduce the accuracy and performance. Most research typically only gives general suggestions for improving security without quantifying the impact, whereby limited practical applicability can be achieved from such recommendations. In addition, many of them have not explained the very crucial issue of processing time in terms of the steps involved in pre-processing, learning, classification, and prediction. Without this, a solution will not be feasible in real-world situations. In turn, our approach tries to fill these gaps by proposing an end-to-end system that emphasizes strict data pre-processing in order to derive from log files high-quality and relevant features that will enhance accuracy and improve predictive performance after proper refinement of the data, and then the application of machine learning algorithms. Moreover, the time complexity analysis for each algorithm allows us to propose an efficient model able to perform the right balance between accuracy and speed of processing. The presence of Apache Spark accelerates data handling and makes the entire workflow smoother, therefore making our solution ideal for big log file analysis in real-time healthcare contexts. The two most important goals-precision and speed-are huge guarantees that our research is one of the great milestones within the field of cybersecurity in general and the elaboration of practical and efficient solutions concerning the identification and mitigation of potentially vulnerable points in healthcare information systems.

III. Big Data Toolkit for Log File Security Analysis

Big Data is a concept that copes with the research, analysis, capture, storage, sharing, and presentation of data. Big Data replaces traditional tools, since those are inefficient in managing and analyzing such large sets of information. Often characterized by the three "Vs": volume, velocity, and variety. Big Data integrates other important aspects such as Value and Veracity [15, 16].

Big Data finds applications across a wide range of industries operating with huge volumes of data every day and usually requires crucial speed [17]. The prominent fields where Big Data can find its central role include Marketing, Monitoring, and Security, among others [18, 19]. In security, Big Data contributes substantially to enhancing the capability that deals particularly in the capture, filtration, and analysis of millions of network events per second, log and audit files included [20]. It provides that third level of protection against cyber-attacks involving:

- First-level security: company security against external attacks [21].
- Second-level security: company protection from vulnerabilities imposed by its users [21].
- Third-level security: impact assessment of the threats detected within infrastructure, which means network traffic back-tracing capability for tracking malware steps and proper action [22].

Big Data analytics tools actually enable cybersecurity experts to analyze a variety of data types from all different sources in real time. These tools are not limited to gathering information but also create correlations and connections [23, 24]. Amongst the major analytical tools for security are:

- Unified Analytics System: Apache Spark, a free software to process large volumes of data, with libraries for different Big Data workloads. Though Spark itself is handy for tasks that require enormous computing power, its application in anomaly detection, both in user behavior and network traffic, will provide information security [25, 26].
- IBM Security QRadar: An enterprise security information and event management product that can collect and analyze log data generated by security systems, network devices, host assets, applications, vulnerabilities, and user activities [27].
- Splunk: it is the premier software platform for search, analysis, and visualization of machine data generated. Splunk allows indexing and real-time analysis of structured, unstructured, and complex logs to enable alerts, event notifications, and reports based on the state of machines [28].
- Apache Metron: An open-source Big Data for security monitoring and analytics that is free. It allows for real-time processing of emitted data; it also has a component for long-term storage. Metron provides a centralized monitoring portal for contextual alerts and attack information [29].

A. Decision Trees

Decision Trees were included because they provide interpretable rule-based classification and can model non-linear interactions among log-derived features [34, 37, 38]. In this study, the tree recursively partitions the feature space according

to the splits that best separate normal and attack traffic, producing decision paths that are easy to inspect. This makes the model attractive for hospital security teams that need understandable alerts, although depth control is required to limit overfitting.

B. K-Nearest Neighbors

K-Nearest Neighbors (KNN) was selected as a strong non-parametric baseline for attack prediction [36, 39]. The classifier assigns a label to each new instance according to the dominant class among its nearest neighbors in feature space, allowing it to capture local and non-linear attack patterns without imposing a fixed functional form. Because KNN is sensitive to feature scaling and the value of k [36], both normalization and hyperparameter tuning were applied before final evaluation. The basic KNN decision principle is illustrated in Fig. 1.

$$Distance(x, x_i) = \sqrt{\sum_{j=1}^{d} \left(x_j - x_{ij}\right)^2}$$

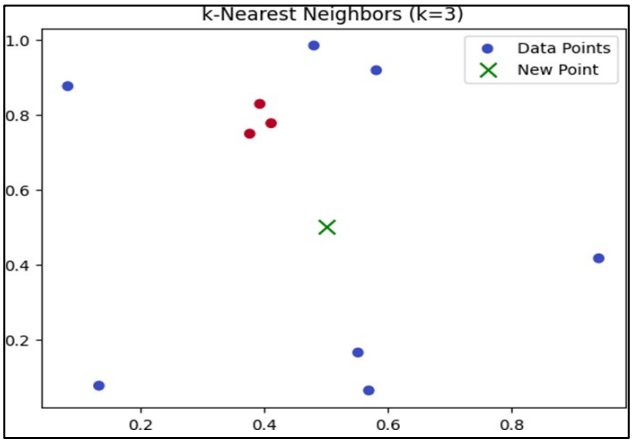


Fig. 1. Basic KNN decision principle.

C. Logistic Regression

Logistic Regression was used as a probabilistic linear baseline for binary intrusion detection [32]. The model estimates the probability that a request belongs to the attack class from a weighted combination of the input features and offers a transparent view of how features contribute to predictions. A conceptual illustration of this decision boundary is shown in Fig. 2.

Its interpretability and low computational cost make it useful in healthcare security analytics, even though purely linear decision boundaries may miss some complex attack relationships. This choice is also consistent with prior work showing that regularized logistic regression can support interpretable anomaly-oriented prediction while remaining suitable for imbalanced security-related settings [32].

$$P(X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d)}}$$

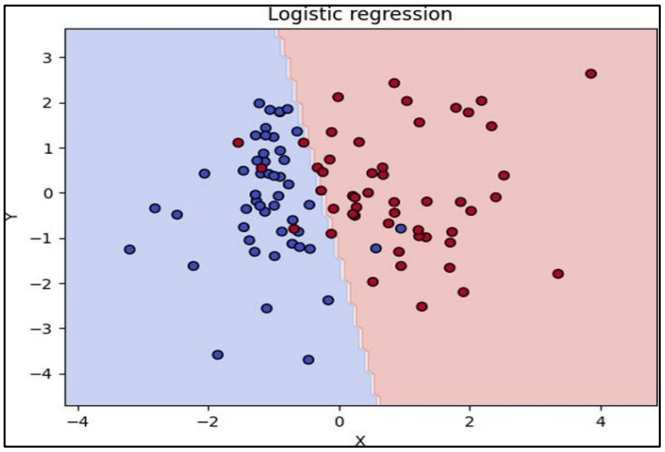


Fig. 2. Illustration of a Logistic Regression model applied to a synthetic binary classification dataset.

D. Comparison Summary

The three classifiers were selected to represent complementary design trade-offs relevant to hospital intrusion detection. Logistic Regression offers transparency and efficiency, Decision Trees provide interpretable non-linear rules, and KNN captures local attack structures with strong discrimination when sufficient labeled examples are available. Comparing them under the same preprocessing and validation protocol allows the study to assess not only accuracy, but also operational fit for large-scale security monitoring.

IV. Design of a Combined Machine Learning and Big Data System for Cyberattack Prediction: Use Case of a University Hospital Information System

The experimental dataset consisted of 1,000,000 labeled HTTP log instances extracted selectively from a university hospital web application environment to form a 30% attack and 70% normal requests, and exported to CSV after preprocessing. The corpus was constructed from raw Apache access logs and request traces, then anonymized to remove patient and user identifiers before feature generation. Labels were assigned as normal or attack using rule-based annotation supported by known malicious payload patterns, status-code anomalies, authentication abuse signatures, and manual verification of representative samples. The retained dataset characteristics are summarized in Table I, while Fig. 3 shows the original structure of the training log data.

TABLE I. Dataset Characteristics

Item	Description
Source	Raw HTTP Log files
Total instances	1,000,000
Class distribution	700,000 normals; 300,000 attack
Retained fields	Packet size, HTTP method, request/response type, URL, status code, parameter count, URL length
Privacy handling	IP addresses and user identifiers are anonymized before modeling

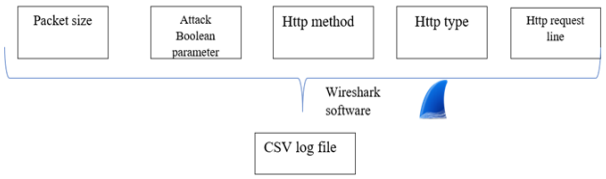


Fig. 3. Structure of the dataset used for training the prediction algorithms

A. Data Preparation and Feature Engineering

Data preparation focused on converting raw web-log events into a consistent feature space suitable for supervised learning. After removing duplicated or incomplete records, categorical fields were encoded, continuous attributes were normalized, and malformed requests were retained only when they contained security-relevant evidence. This step reduced noise while preserving the behavioral signals needed for attack prediction.

The final feature set included HTTP method, request/response type, packet size, request URL, status code, URL length, and parameter count. These features were chosen because they capture request structure and protocol behavior in a form that remains compatible with the information retained from the original server logs.

B. Predictive Modeling and Experimental Design

Three supervised classifiers were trained on the same labeled dataset: KNN, Logistic Regression, and Decision Trees. The data were partitioned with a stratified 80/20 train-test split so that the original normal/attack ratio was preserved in both subsets. All model selection steps were performed only on the training data. The overall workflow for model construction, application, and evaluation is illustrated in Fig. 4.

- Hyperparameter tuning was conducted using grid search with five-fold stratified cross-validation. For KNN, we searched k in {3, 5, 7, 9, 11}, weight schemes {uniform, distance}, and distance metrics {euclidean, manhattan}. For Decision Trees, we explored the split criterion {gini, entropy}, max depth {3, 5, 10, None}, and min samples split {2, 5, 10}. For Logistic Regression, we evaluated C in {0.01, 0.1, 1, 10} with L2 regularization using the liblinear and lbfgs solvers.
- Model performance was reported on the held-out test set using accuracy, precision, recall, F1-score, false positive rate, and processing time. Apache Spark was used to accelerate preprocessing and batch inference at larger scales, enabling a consistent comparison between traditional and distributed execution. The best configurations selected by cross-validation were KNN (k = 5, distance weighting, Euclidean distance), Decision Tree (gini, max_depth = 10, min_samples_split = 5), and Logistic Regression (C = 1, L2, liblinear). The selected hyperparameter values are summarized in Table II.

TABLE II. SELECTED HYPERPARAMETER VALUES

Model	Selected configuration
KNN	k = 5; distance weighting; Euclidean metric
Decision Tree	gini; max_depth = 10; min_samples_split = 5
Logistic Regression	C = 1; L2 penalty; liblinear solver

C. Results Analysis and Discussion

The evaluation examined both predictive quality and deployment-oriented efficiency. Final classification metrics were computed on a held-out test set of 200,000 instances, while processing-time measurements were recorded across dataset sizes from 100,000 to 1,000,000 instances to assess scalability under increasing log volume.

D. Classification Performance and Operational Metrics

KNN achieved the strongest overall results on the held-out test set, reaching 95.66% accuracy and substantially outperforming Logistic Regression (85.30%) and Decision Trees (83.20%). Its advantage was not limited to accuracy: KNN also produced the highest precision and F1-score, indicating that it balanced correct attack detection with reliable alert generation. The full metric breakdown is reported in Table III, and the cross-size accuracy trend is shown in Fig. 5.

From an operational perspective, the false positive rate is especially important for hospital environments because excessive false alarms can overwhelm security teams and reduce trust in automated monitoring. KNN yielded the lowest false positive rate (3.60%), whereas Logistic Regression and Decision Trees produced 8.00% and 9.00%, respectively. This result strengthens the case for KNN as the most deployment-ready classifier among the models evaluated.

TABLE III. CLASSIFICATION PERFORMANCE

Model	Accuracy	Precision	Recall	F1-score	FPR	Specificity
KNN	95.66%	91.79%	93.93 %	92.85 %	3.60 %	96.40%
Logistic Regression	85.30%	78.87%	69.67 %	73.98 %	8.00 %	92.00%
Decision Tree	83.20%	75.58%	65.00 %	69.89 %	9.00 %	91.00%

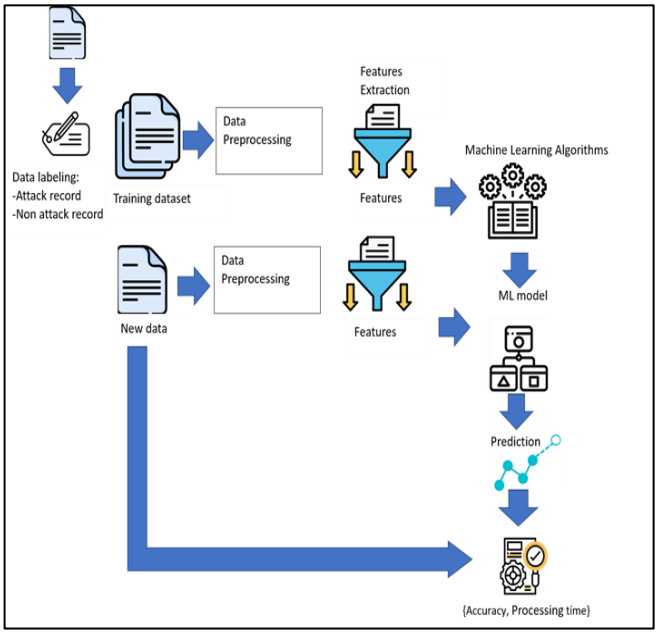


Fig. 4. Process of constructing, applying, and evaluating the predictive model for HTTP attacks prediction.

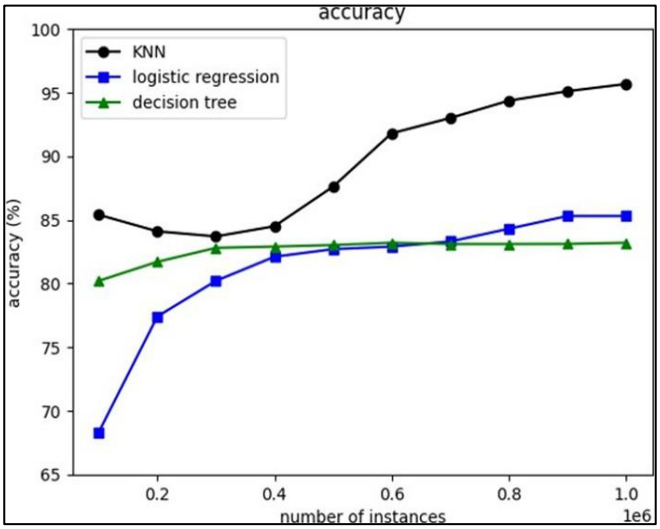


Fig. 5. Accuracy comparison of machine learning algorithms (KNN, logistic regression, and decision tree) for HTTP attack prediction across different dataset sizes.

E. Processing Time Analysis

Processing-time analysis confirmed the value of distributed execution for large-scale hospital log analysis. Across all three models, Spark reduced runtime as dataset size increased, with the strongest absolute gains observed for KNN and Decision Trees at the 1,000,000-instance scale. Although Logistic Regression remained the fastest classifier under Spark (6.0 s), KNN matched Decision Tree at 6.5 s while preserving a much stronger detection profile, making it the best accuracy-efficiency compromise in the study. Table IV summarizes the detailed timing values, and Fig. 6 highlights the corresponding processing-time trend.

TABLE IV. PROCESSING TIME COMPARISON (SECONDS)

Dataset size	Model	Without Spark (s)	With Spark (s)
100,000	KNN	1.8	1.1
100,000	Logistic Regression	1.4	1.2
100,000	Decision Tree	2.0	1.4
300,000	KNN	4.5	2.6
300,000	Logistic Regression	3.0	2.8
300,000	Decision Tree	4.7	2.9
500,000	KNN	6.8	3.7
500,000	Logistic Regression	4.2	3.9
500,000	Decision Tree	7.3	4.1
700,000	KNN	8.9	4.8
700,000	Logistic Regression	5.1	4.9
700,000	Decision Tree	9.6	5.0
1,000,000	KNN	12.0	6.5
1,000,000	Logistic Regression	6.5	6.0
1,000,000	Decision Tree	12.6	6.5

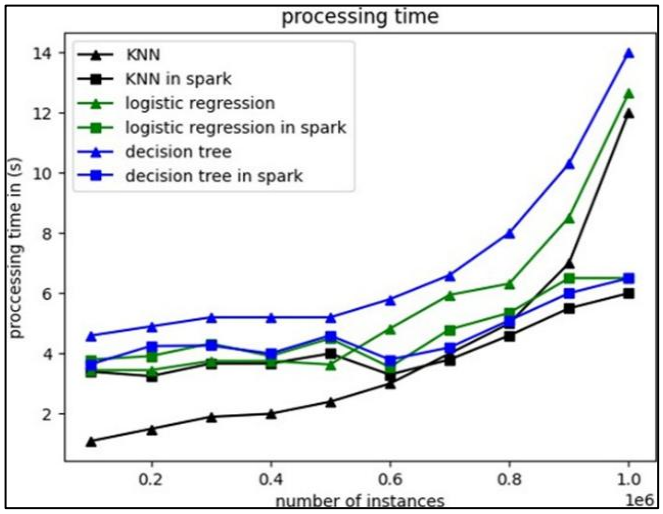


Fig. 6. Processing time comparison of KNN, logistic regression, and decision tree algorithms, both with traditional processing and using Spark.

V. CONCLUSION

This study presented a coherent framework for cyberattack prediction in e-Health information systems by combining structured log analysis, supervised machine learning, and Apache Spark-based distributed processing. Using a labeled corpus of 1,000,000 HTTP log instances from a university hospital context, the study demonstrated that rigorous preprocessing and feature engineering can support reliable detection of malicious web activity against healthcare services.

Among the evaluated classifiers, KNN delivered the best overall performance, achieving 95.66% accuracy, 91.79% precision, 93.93% recall, a 92.85% F1-score, and the lowest false positive rate (3.60%). These results are especially important for operational deployment because they indicate a lower alert burden on hospital security teams compared with Logistic Regression and Decision Trees.

The processing-time experiments further showed that Apache Spark materially improves scalability, cutting KNN runtime from 12.0 s to 6.5 s at 1,000,000 instances and enabling practical large-scale analysis without sacrificing detection quality. Taken together, the results suggest that KNN plus distributed execution offers the most favorable balance of accuracy, responsiveness, and operational usability for intrusion monitoring in e-Health environments.

Future work can extend this framework toward multi-class attack labeling, streaming analytics for continuous monitoring, and external validation on logs collected from additional hospital systems. These steps would further strengthen generalizability while preserving the deployment-oriented focus established in the present study.

REFERENCES

[1] Aceto G, Persico V, Pescapè A (2018) The role of information and communication technologies in healthcare: taxonomies, perspectives, and challenges. J Netw Comput Appl 107:125–154

[2] Husák M, Komárková J, Bou-Harb E, Čeleda P (2018) Survey of attack projection, prediction, and forecasting in cyber security. IEEE Commun Surv Tutor 21(1):640–660.

[3] OWASP (2024) OWASP top ten web application security risks - 2024. Available from: https://owasp.org/www-project-top-ten/

[4] Gzar DA, Mahmood AM, Abbas MK (2022) A comparative study of regression machine learning algorithms: tradeoff between accuracy and computational complexity. Math Model Eng Prob 9(5):1379–1384

[5] Abdalla RR, Jumaa AK (2022) Log file analysis based on machine learning: a survey. UHD J Sci Technol 6(2):77–84

[6] Hubballi N, Khandait P (2023) Event log analysis and correlation: a digital forensic perspective. In: Artificial Intelligence (AI) in Forensic Sciences, p 195

[7] Sittig DF, Wright A (2023) A guide to mitigating audit log-related risk in medical professional liability cases. J Healthc Risk Manag

[8] Prabha P, Kumar S, Shree N, Sundaram R, Nishanthi HM, Pranesh D (2024) Cybersecurity in healthcare: safeguarding patient data

[9] Wickramage C, Sahama T, Fidge CJ (2016) Anatomy of log files: implications for information accountability measures

[10] Landauer M, Wurzenberger M, Skopik F et al (2018) Dynamic log file analysis: an unsupervised cluster evolution approach for anomaly detection. Comput Secur 79:116

[11] Zhong M, Zhou Y, Chen G (2021) A security log analysis scheme using deep learning algorithm for IDSS in social network. Secur Commun Netw 2021:5542543

[12] Skopik F, Wurzenberger M, Landauer M (2022) Detecting unknown cyber security attacks through system behavior analysis. In: Cybersecurity of Digital Service Chains: Challenges, Methodologies, and Tools, Springer, Cham, p 103–119

[13] Azizi Y, Azizi M, Elboukhari M (2019) Log files analysis using MapReduce to improve security. Procedia Comput Sci 148:37–44

[14] Jeon D, Tak B (2022) Blackeye: automatic IP blacklisting using machine learning from security logs. Wireless Netw 28(2):937–948

[15] Khan N, Yaqoob I, Hashem IAT et al (2014) Big data: survey, technologies, opportunities, and challenges. Sci World J 2014:712826

[16] Bhadani AK, Jothimani D (2016) Big data: challenges, opportunities, and realities. Effective Big Data Manage Opp Implement 1–24

[17] Katal A, Wazid M, Goudar RH (2013) Big data: issues, challenges, tools and good practices. In: Sixth Int Conf Contemp Comput (IC3), IEEE, p 404–409

[18] Bello-Orgaz G, Jung JJ, Camacho D (2016) Social big data: recent achievements and new challenges. Inf Fusion 28:45–59

[19] Sagiroglu S, Sinanc D (2013) Big data: a review. In: Int Conf Collabor Technol Syst (CTS), IEEE, p 42–47

[20] Amanullah MA, Ariyaluran RAH, Nasaruddin FH et al (2020) Deep learning and big data technologies for IoT security. Comput Commun 151:495–517

[21] Kwan P, Ho CJ (2003) Multiple tiered network security system, method and apparatus

[22] Griffin M (2013) Assessment of run-time malware detection through critical function hooking and process introspection against real-world attacks. The University of Texas at San Antonio

[23] Rassam MA, Maarof M, Zainal A et al (2017) Big data analytics adoption for cybersecurity: a review of current solutions, requirements, challenges and trends. J Inf Assur Secur 12(4)

[24] Mahdavi P (2014) Google correlations: new approaches to collecting data for statistical network analysis. University of California, Los Angeles

[25] Arif Z, Zeebaree SRM (2024) Distributed systems for data-intensive computing in cloud environments: a review of big data analytics and data management. Indones J Comput Sci 13(2)

[26] Saeed MA, Saeed MA (2024) Real-time diabetes detection using machine learning and Apache Spark. In: 4th Int Conf Emerg Smart Technol Appl (eSmarTA), IEEE, p 1–6

[27] Gnatyuk S, Berdibayev R, Aleksander M et al (2024) Software system for cybersecurity events correlation and incident management in critical infrastructure. In: Data-Centric Business and Applications: Advancements in Information and Knowledge Management, Springer, Cham, p 247–269

[28] Razak AA, Ruzaili HH, Zolkifli MF (2024) Study on machine learning implementation in cybersecurity for security defend and attack. Borneo Int J 7(2):27–40

[29] Franzén MF, Tyrén N (2021) Anomaly detection for automated security log analysis: comparison of existing techniques and tools

[30] Langone R, Cuzzocrea A, Skantzos N (2020) Interpretable anomaly prediction: predicting anomalous behavior in Industry 4.0 settings via regularized logistic regression tools. Data Knowl Eng 130:101850

[31] Huang C (2018) Featured anomaly detection methods and applications. University of Exeter, UK

[32] Mahbooba B, Timilsina M, Sahal R, Serrano M (2021) Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. Complexity 2021:6634811

[33] Liu C, Li B, Vorobeychik Y, Oprea A (2017) Robust linear regression against training data poisoning. In: Proc 10th ACM Workshop Artif Intell Secur, p 91–102

[34] Halder RK, Uddin MN, Uddin MA, Aryal S, Khraisat A (2024) Enhancing k-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. J Big Data 11(1):113

[35] Azam Z, Islam MM, Huda MN (2023) Comparative analysis of intrusion detection systems and machine learning based model analysis through decision tree. IEEE Access

[36] Lakshminarasimman S, Ruswin S, Sundarakantham K (2017) Detecting DDoS attacks using decision tree algorithm. In: 4th Int Conf Signal Process, Commun Netw (ICSCN),IEEE, p 1–6

[37] Yusof AR, Udzir NI, Selamat A (2016) An evaluation on KNN-SVM algorithm for detection and prediction of DDoS attack. In: Trends Appl Knowl-Based Syst Data Sci: 29th Int Conf Ind Eng Other Appl Appl Intell Syst (IEA/AIE 2016), Springer, Morioka, Japan, p 95–102

[38] Sannigrahi M, Thandeeswaran R (2024) Predictive analysis of network based attacks by hybrid machine learning algorithms utilizing Bayesian optimization, logistic regression and random forest algorithm. IEEE Access

[39] Chen JJ, Tsai C-A, Moon H, Ahn H, Young JJ, Chen C-H (2006) Decision threshold adjustment in class prediction. SAR QSAR Environ Res 17(3):337–352

Solar Irradiance Forecasting Approaches Based on Machine Learning: A Systematic Literature Review

Sempe Thom Leholo[1]*, Chunling Tu[2], Topside Ehleketani Mathonsi[3]
Dept of Computer Systems Engineering, Tshwane University of Technology, Pretoria, South Africa[1, 2]
Dept of Information Technology, Tshwane University of Technology, Pretoria, South Africa[3]

***Abstract*—The prediction of solar irradiance plays a crucial role in the design, performance, and stability of renewable energy sources, and especially photovoltaic (PV) power generation. Accurate forecasting helps in managing energy, grid stability, and integration of solar energy in contemporary power systems. The study is a Systematic Literature Review (SLR) of 37 recent (2019-2025) peer-reviewed papers on solar irradiance forecasting that apply Machine Learning (ML), Deep Learning (DL), and hybrid or ensemble modelling methods. The review is based on the Preferred Reporting Items of Systematic Reviews and Meta-Analyses (PRISMA 2020) to make it transparent and reproducible. A thorough search of seven large databases, such as Google Scholar, IEEE Xplore, Web of Science, Springer Nature Link, ScienceDirect, MDPI and the ACM Digital Library, was conducted to find relevant studies. Based on a structured synthesis of the chosen literature, the findings suggest that there is a definite methodological change in the traditional ML methods to DL and hybrid modelling structures. Although the classical ML algorithms have low computational complexity and can be effectively used to make short-term predictions, DL architectures consistently outperform them in terms of capturing nonlinear temporal and spatial patterns in solar irradiance data. Moreover, hybrid models combining DL architectures with signal decomposition and feature fusion methods also improve predictive accuracy. Nevertheless, the review notes that there are a number of ongoing shortcomings, such as the lack of geographic generalizability because of single-site dominance, the lack of consistency in reporting computational efficiency, the inconsistency of evaluation metrics, the lack of robustness testing in dynamic weather conditions, and a strong bias towards short-term forecasting horizons. In order to fill these gaps, future studies need to focus on multi-site and cross-climatic validation, domain adaptation using transfer learning, designing lightweight models to deploy in real-time, standardised benchmarking guidelines, and broaden their scope to medium and long-term forecasting with enriched meteorological inputs. Overall, the results offer an evidence-based, systematic review of existing trends in methodology and emphasise the need to balance predictive accuracy with generalizability, efficiency, and practical application in solar energy forecasting systems.**

Keywords*—Solar irradiance forecasting; Machine Learning; Deep Learning; hybrid models; Systematic Literature Review

I. Introduction

Solar power is a major renewable resource that can be used to reduce climate change and to supply the ever-growing global electricity demand. However, solar irradiance is very temporal and sensitive to meteorological conditions like cloud dynamics, atmospheric composition, temperature and humidity making it difficult to make precise predictions [1]. This natural intermittency presents an element of uncertainty in the operation of Photovoltaic (PV) power systems and electrical grids to the stability, reliability and economic performance of the system [2]. The inaccuracy of the forecasts can lead to power imbalances, poor dispatch choices, reserve overage, and operational expenses, which restrict the large-scale integration of solar energy [3]. Therefore, improving the accuracy and strength of solar irradiance forecasting is a focal point of research in the field of renewable energy and grid integration [4].

To address these issues, various forecasting strategies have been explored, including classical statistical models and physical models as well as data-driven Machine Learning (ML) models [5]. Though conventional models are based on pre-established assumptions and simplified physical relationships, they do not usually describe the nonlinear and stochastic interactions between solar irradiance and atmospheric variables [6]. ML-based methods have thus become popular because of their ability to acquire complicated patterns using historical irradiance and meteorological information. Earlier experiments have used regression models, tree-based algorithms, Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), and Deep Learning (DL) models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks to enhance the accuracy of forecasts. Ensemble and hybrid models have also been put forward in order to capitalise on complementary model strengths and to enhance performance under different climatic conditions [7].

Although they are said to be successful, in the real world of operation, ML-based forecasting models are usually limited. Issues with data quality, non-stationarity and distributional changes due to seasonal variability, climate anomalies, or sensor degradation can cause Model performance to degrade [8]. In addition, most studies based on site-specific data and controlled experimental design limit model extrapolation across geographical areas and time spans. More advanced methods like feature selection, hyperparameter optimisation and probabilistic forecasting have been proposed to enhance robustness and reliability. However, the quantification of uncertainty and operational resilience has not been investigated thoroughly in most of the literature [9, 10]. New methods such as transfer learning and federated learning have demonstrated the potential to deal with data scarcity, privacy issues and regional heterogeneity. Conversely, Explainable Artificial Intelligence (XAI) methods are becoming more popular to enhance the

*Corresponding author

interpretability of models and enable trust in ML-driven prediction systems [11-13].

This Systematic Literature Review (SLR) critically analyses and synthesises recent advances in ML, DL and hybrid-based solar irradiance forecasting reported between 2019 and 2025. The review systematically finds and categorises available forecasting methods into ML, DL and hybrid architectures and compares their reported performance based on popular metrics. In addition to performance comparisons, the study will discuss the advantages and disadvantages of these techniques in practical implementation. The discussion identifies some common problems, such as inadequate generalisation of models across geographic and climatic boundaries, preponderance of short-term forecasting research, underreporting of operational and computational needs and weak performance in highly variable weather. Lastly, the review identifies several significant research gaps and future research directions. It also offers practical insights and methodological advice on developing reliable, scalable and application-oriented solar irradiance forecasting models for renewable energy systems.

This SLR makes the following contributions to the literature:

- Provides a comprehensive systematic review of ML, DL and hybrid methods for solar irradiance forecasting.
- Provide a structured classification of forecasting approaches into traditional ML, DL and hybrid/ensemble models used for solar irradiance prediction.
- Provide a comparative evaluation of forecasting performance across studies using commonly reported metrics such as RMSE, MAE, nRMSE and the coefficient of determination (R^2).
- Provide a critical analysis of ongoing limitations such as limited model generalizability across climatic regions, insufficient reporting of operational and computational metrics, the dominance of short-term forecasting studies and limited evaluation under highly variable weather conditions.
- Identification of key research opportunities such as the development of generalizable multi-site forecasting models, integration of NWP with DL for longer forecasting horizons, lightweight and efficient architectures for real-time deployment, probabilistic forecasting frameworks and the creation of standardised benchmark datasets.

The rest of this study will be organised in the following way: Section II will provide the background. Section III describes the methodology of the review according to the PRISMA framework. Section IV introduces the chosen studies to review and provides a comparative analysis of the forecasting models and their performance. Section V explains the major issues of research, limitations, and the practical deployment. Section VI points out the future research directions and methodological recommendations. Lastly, Section VII gives the limitations of this review and the conclusion of the work.

II. Background

In order to contextualise the existing studies on solar irradiance forecasting, it is necessary to develop a clear vision of the inherent features of the variability of solar irradiance, the constraints of the traditional forecasting methods, and the increasing role of ML in renewable energy systems. This section examines these background concepts and systematically prepares the ground on which the detailed analysis in the following sections will be based.

A. Solar Irradiance Variability

The other important parameter that influences PV power production is solar irradiance. But solar irradiance is highly variable since it is affected greatly by the atmospheric and meteorological conditions [14]. Variables such as the concentration of aerosols, movement of clouds, humidity, temperatures, and geometry of the sun contribute to the variability of solar irradiance [15]. Fluctuation of solar irradiance presents a number of challenges to the management of power systems, particularly in the increased use of solar energy in modern electrical systems [15]. The issue is that, contrary to the traditional fuels, solar energy cannot be turned on and off according to the load needs [16]. Thus, it is important to forecast the amount of solar irradiance accurately to ensure stability and proper power load management, as well as participation in energy market activities. Short-term variability may bring about voltage fluctuations and frequency instability, whereas longer-term forecasting errors can be inefficient scheduling of backup generation and higher operating costs [17]. Consequently, the nature and cause of the variability of solar irradiance is the key area of concern in renewable energy studies.

B. Challenges in Solar Irradiance Forecasting

The prediction of solar irradiance remains a challenge because it is stochastic, nonlinear and dynamic in nature [18]. Although the radiation of the sun itself is deterministic, it interacts with a number of atmospheric variables, including clouds, aerosols, and water vapour, which make it stochastic [19]. Solar irradiance forecasting at the beginning was largely grounded on statistical and physics-based modelling with the use of Numerical Weather Predictions (NWP) [20]. Although these techniques are promising when studied at a greater time scale, they perform poorly in forecasting short-term variations in the solar irradiance [21]. The reason is that they are not effective in forecasting irradiance in partly cloudy skies. Moreover, the availability of data is another issue, as not all the data may have been available and there may be noise in sensors that capture irradiance measurements [22]. The lack of data and climatological differences between various geographical locations also make the predictive models training and deployment more difficult [23]. Such problems have inspired the creation of data-driven methods that can directly acquire complex patterns through historical observations.

C. Machine Learning for Solar Irradiance Forecasting

Machine Learning (ML) algorithms have proven to be quite effective in forecasting the performance of solar energy systems

thanks to their ability to deal with non-linear relations and build complicated representations of varied inputs [24]. The ML algorithms have proven to be the best approaches in solar radiation forecasting in the short term as well as in day-ahead predictions [25]. This has been attributed to the fact that the algorithms have the capability to simulate very complex interactions of non-linear meteorological variables such as the solar irradiation, temperature, wind speed, and humidity [26]. ML algorithms do not require any assumption regarding physical processes, as opposed to other traditional methods, which require some physical properties. The algorithms are capable of making good use of historical irradiation and meteorological data to enhance their forecast effectiveness [27]. Solar irradiance forecasting has been performed with different ML methods, such as regression-based methods, Decision Trees (DTs), Support Vector Machines (SVMs), ensemble learning methods, such as Random Forests (RFs) and Gradient Boosting (GB), Deep Learning (DL), and hybrid methods that integrate data-driven methods with other methods [28, 29]. The techniques are highly appreciated due to their flexibility in modelling. The computation needs, however, differ significantly among model classes.

ML algorithms that are applied in predicting solar irradiance have three major classifications. They are unsupervised, hybrid and supervised [30]. Among all these types, the supervised ML algorithms have demonstrated excellent performance in the case of sufficiently good quality training data [31]. These algorithms mostly use handcrafted features extracted from historical solar irradiance data, weather parameters, and time indicators to obtain accurate predictions [32]. Nevertheless, their effectiveness is highly dependent on having plenty of data, being vulnerable to overfitting, and low generalisation abilities for unknown locations with different climates [33]. In order to overcome these issues, the DL algorithms have gained a great deal of interest in recent years due to their ability to directly obtain complex features based on raw data [34]. The most popular DL structures to use in modelling solar irradiance time series include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory Networks (LSTM) [35, 36]. Particularly, LSTM algorithms have been effective in learning long-term dependencies in solar irradiance variations [37, 38].

Unlike supervised learning methods, unsupervised learning methods do not need labelled data and are mainly applied to the task of clustering, pattern discovery, and regime identification [39]. These methods are very useful in understanding the underlying structure of irradiance and meteorological data and can be used to aid forecasting by adaptive model selection or feature extraction [40]. But in contrast to the supervised methods, the unsupervised methods do not directly produce irradiance forecasts and are generally only feasible when combined with predictive models [41].

Hybrid forecasting methods are also increasingly being combined with supervised and unsupervised learning techniques using statistical or physical models. These kinds of hybrid models take advantage of complementary capabilities of different algorithms to enhance the accuracy and reliability of predictions in different weather conditions [42]. Hybrid Machine Learning models, especially, have gained quite a lot of attention in relation to predicting solar irradiance due to the nonlinear, non-stationary, and stochastic nature of solar energy [43]. The solar irradiance is also subject to a number of meteorological conditions, such as cloud cover, temperature, humidity, wind speed, and aerosol content; such factors cannot be well modelled using only one modelling method [44]. Hybrid models are able to give a more realistic account of these complex dynamics by combining complementary learning approaches.

Even though hybrid approaches tend to enhance predictive reliability, they can add complexity to the model, raise the cost of computation, and raise the design and tuning needs [45]. Hybrid ML models, in general, have better accuracy, flexibility, and generalizability of multiple forecasting horizons [46]. They are especially fit in real-world solar energy systems because they are able to combine spatial, temporal, and physical information, and accurate forecasting is important in grid stability, energy management, and to enable the successful integration of renewable energy resources [47].

Though significant progress has been made in this area, there is a tendency for ML, DL, and ML-DL forecasting algorithms to be ineffective in practice because of non-stationary processes, seasonality changes, and changes in the distribution of weather patterns [48]. In a bid to mitigate these challenges, ensemble learning techniques, probabilistic forecasting and hyperparameter optimisation approaches have been proposed. However, there are inconsistent uptakes of these methods in the literature. These difficulties prove the necessity of the systematic analysis and critical synthesis of the existing approaches.

This development of ML-based solar irradiance forecasting studies forms the basis of the SLR in this study, which seeks to synthesise existing knowledge, evaluate the methodological strengths and limitations, and uncover gaps and future research directions in the field of ML-based solar irradiance forecasting.

III. Methodology

The methodology of this SLR defines the structured approach adopted to identify, screen, and synthesise existing research on ML and DL-based solar irradiance forecasting. The review was conducted in accordance with the PRISMA 2020 guidelines to ensure transparency, methodological rigour, and reproducibility.

A. Review Protocol and Prisma Guidelines

The SLR in this study was conducted in accordance with the PRISMA framework, which provides a structured process for identifying, screening and selecting relevant studies. A predefined review protocol was established to guide the search strategy, inclusion and exclusion criteria, data extraction and synthesis process. This protocol was designed to minimise selection bias and ensure consistency throughout the review.

B. Research Questions

The SLR in this study was guided by the following Research Questions (RQs):

- RQ1: What types of predictive models are used for solar irradiance forecasting and what model architectures are commonly adopted?

- RQ2: What data sources, input features, feature selection methods and feature engineering techniques are utilised in these forecasting models?
- RQ3: What forecasting horizons are primarily addressed in the literature?
- RQ4: What model optimisation and tuning techniques, evaluation metrics, and validation methods are employed to assess forecasting performance?
- RQ5: What key findings, limitations, challenges and research gaps are reported in the literature regarding the development, generalizability and real-world deployment of solar irradiance forecasting models?

C. Research Objectives

To address these research questions, the following Research Objectives (ROs) were formulated:

- RO-1: To identify and analyse the types of predictive models, including ML, DL and hybrid approaches, used for solar irradiance forecasting and to examine their commonly adopted architectures.
- RO-2: To investigate the data sources, input features, feature selection methods and feature engineering techniques utilised in solar irradiance forecasting models.
- RO-3: To examine the forecasting horizons addressed in the literature, including short-term, medium-term and long-term solar irradiance prediction.
- RO-4: To analyse the optimisation and tuning techniques, evaluation metrics and validation methods used to assess and compare forecasting model performance.
- RO-5: To identify the key findings, limitations, challenges and research gaps reported in the literature regarding the generalizability and real-world deployment of solar irradiance forecasting models.

D. Search Strategy

A comprehensive search strategy was developed using combinations of relevant keywords and Boolean operators. These keywords were combined using logical operators (AND, OR) to retrieve relevant studies. The search was limited to peer-reviewed publications to ensure scientific quality. The primary search terms included:

("solar irradiance forecasting" OR "solar radiation prediction") AND

("machine learning" OR "deep learning" OR LSTM OR CNN OR GRU OR transformer OR "ensemble learning" OR "hybrid models" OR "sky images")

The literature search was limited to studies published between 2019 and 2025 to capture recent advances in ML and DL techniques for solar irradiance forecasting.

E. Data Source Material

A comprehensive literature search was conducted using seven major electronic databases widely recognised in the fields of renewable energy, power systems and artificial intelligence. These databases include Google Scholar, IEEE Xplore, Web of Science, SpringerLink, ScienceDirect, MDPI and ACM Digital Library. These databases were selected to ensure broad coverage of high-quality, peer-reviewed research.

F. Inclusion and Exclusion Criteria

Clear inclusion and exclusion criteria were established to maintain the rigour and relevance of the SLR. The journal or conference articles were included under the following strict conditions:

- They have been published between 2019 and 2025.
- They have been peer-reviewed and are accredited according to the latest Department of Higher Education and Training-accredited journal list.
- They focused on ML and DL based solar irradiance or solar radiation forecasting.
- They provided an empirical evaluation using quantitative performance metrics.

Studies were excluded if:

- They have been published before 2019.
- They have no empirical or experimental validation.
- The metric evaluation results are not explicitly discussed.
- They are written in other languages instead of English.
- They are tutorials, letters, editorials, review articles, opinion papers, and blogs.

These criteria ensured that the review captured recent, evidence-based research while excluding non-technical or outdated sources, thereby strengthening the reliability and relevance of the synthesised findings.

G. Study Selection Process

The study selection followed the PRISMA four-stage process:

- Identification: All records retrieved from databases were compiled, and duplicates were removed.
- Screening: Titles and abstracts were screened to exclude irrelevant studies.
- Eligibility: Full-text articles were assessed based on inclusion and exclusion criteria.
- Inclusion: Studies meeting all criteria were included in the final review.

A comprehensive literature search was conducted across relevant academic databases to identify studies on solar irradiance forecasting using ML techniques. A total of 162 records were initially identified through database searching.

Following identification, 66 duplicate records were removed, leaving 92 unique records for screening. The titles and abstracts of these records were screened to assess relevance to the research objectives. During this stage, 43 records were excluded for being irrelevant to solar irradiance forecasting or for lacking ML or DL methodologies.

The remaining 49 reports were sought for full-text retrieval; all were successfully obtained. These reports were then assessed for eligibility based on predefined inclusion and exclusion criteria. As a result of the full-text assessment, 12 reports were excluded for reasons including a lack of empirical or experimental validation and for not explicitly discussing metric evaluation results.

Ultimately, 37 studies met all inclusion criteria and were included in the final review. Each included study corresponded to a single report. The complete study selection process is illustrated in the PRISMA 2020 flow diagram shown in Fig. 1.

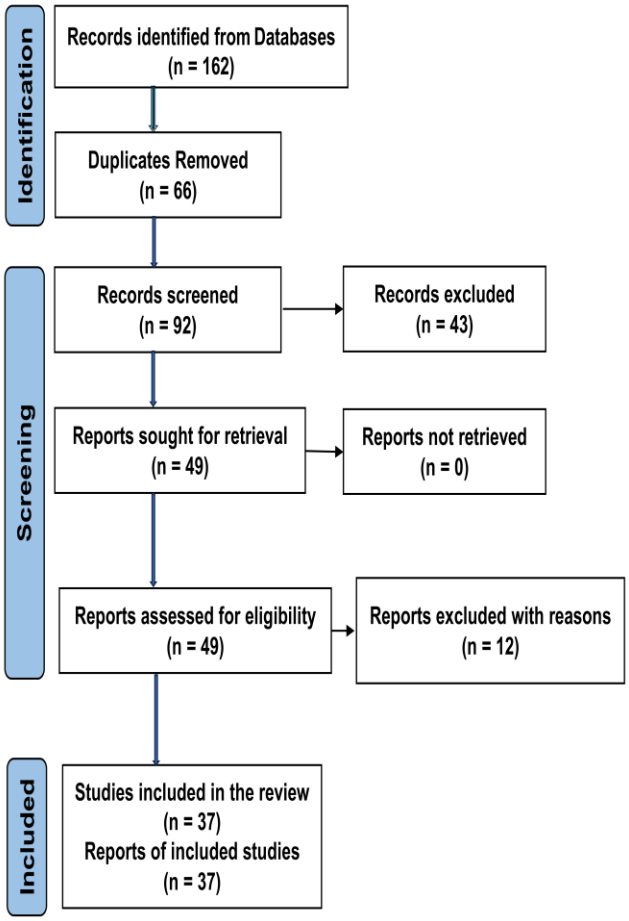


Fig. 1. PRISMA 2020 flow diagram of this study.

This systematic selection process ensures that only high-quality, relevant, and methodologically sound studies contribute to this review's findings.

H. Quality Assessment

The quality of the included studies was systematically assessed to determine the reliability and methodological rigour of the research in a simplified checklist-based framework. Instead of numeric scoring, the studies were graded according to four broad criteria that encompassed the quality of data, methodological rigour, the strategy of validation and clarity of reporting. Each of the criteria was assessed based on three possible outcomes, i.e., Yes, Partial or No (or not reported). The rating of "Yes" means that the criterion is completely met, the rating of "Partial" means that the criterion is met partially, and the rating of "No" means that the criterion is not met or not mentioned.

The following defines the assessment criteria: data quality evaluates whether the size, duration and source of the data are clearly described and adequate to the forecasting task; methodological rigour evaluates whether the model architecture, hyperparameter tuning and methodological decisions are clearly explained and reproducible; validation strategy evaluates whether suitable validation methods, such as time-based splitting or cross-validation, are employed to avoid data leakage and guarantee robustness; reporting clarity evaluates A Partial rating is also used to prevent subjectivity when a criterion is covered, but not detailed or comprehensive enough. Examples include a mention of a dataset source, but its size or duration is not completely stated, or validation is mentioned, but an unsuitable or imprecisely described method is employed.

Based on these criteria, each study was classified using a rule-based approach where high quality means that all four criteria were rated as "Yes", medium quality means that at least two criteria were rated as "Yes" and no criterion was rated as "No" (i.e., only "Partial" allowed otherwise), and low quality means any criterion was rated as "No". The final synthesis has filtered out studies that are of low quality in order to have a review that is grounded on solid and methodologically sound evidence. Table I includes the quality assessment criteria and checklist, and Table II includes the summary of the assessment results of all included studies. The findings show that none of the chosen studies were of low quality according to the established criteria; all of them are of high and middle quality. This implies that all of the included studies have a minimum level of methodological rigour and reporting adequacy. The most frequent restrictions found in medium-quality studies are associated with missing reporting of validation processes and transparency of hyperparameter tuning. This systematic but streamlined evaluation system enhances transparency, minimises subjectivity, and provides uniformity in assessing the quality of the methodology of the reviewed literature.

TABLE I. QUALITY ASSESSMENT CRITERIA AND CHECKLIST

Criterion	Description	Yes	Partial	No
Data Quality	Dataset source, size, duration, and type are clearly described and sufficient for the forecasting task	Clearly specified dataset (source, duration, resolution, inputs)	Partially described dataset (missing size/duration/type details)	Datasets unclear or not reported

Methodological Rigour	Model architecture, feature engineering, and hyperparameter tuning are clearly justified and reproducible	Full model description + tuning/justification provided	Model described but limited tuning/justification	Model unclear or not reproducible
Validation Strategy	Use of appropriate validation (chronological split, CV, walk-forward), preventing leakage	Clearly defined robust validation method	Validation mentioned but unclear/incomplete	No validation strategy reported
Reporting Clarity	Metrics clearly reported (RMSE, MAE, R^2, etc.) with transparency	Clear metrics and results reported	Metrics reported but incomplete/inconsistent	Metrics unclear or missing

TABLE II. QUALITY ASSESSMENT RESULTS FOR INCLUDED STUDIES

Ref	Data Quality	Methodological Rigour	Validation Strategy	Reporting Clarity	Quality
[49]	Yes	Yes	Yes	Yes	High
[50]	Yes	Yes	Yes	Yes	High
[51]	Yes	Yes	Yes	Yes	High
[52]	Partial	Yes	Partial	Yes	Medium
[53]	Partial	Yes	Partial	Yes	Medium
[54]	Yes	Yes	Yes	Yes	High
[55]	Yes	Yes	Yes	Yes	High
[56]	Yes	Yes	Yes	Yes	High
[57]	Yes	Yes	Partial	Yes	Medium
[58]	Yes	Yes	Yes	Yes	High
[59]	Yes	Yes	Yes	Yes	High
[60]	Yes	Yes	Partial	Yes	Medium
[61]	Yes	Yes	Yes	Yes	High
[62]	Yes	Yes	Yes	Yes	High
[63]	Yes	Yes	Yes	Yes	High
[64]	Partial	Yes	Partial	Yes	Medium
[65]	Yes	Yes	Yes	Yes	High
[66]	Yes	Yes	Yes	Yes	High
[67]	Yes	Yes	Yes	Yes	High
[68]	Yes	Yes	Partial	Yes	Medium
[69]	Yes	Yes	Yes	Yes	High
[70]	Yes	Yes	Yes	Yes	High
[71]	Yes	Yes	Yes	Yes	High
[72]	Yes	Yes	Yes	Yes	High
[73]	Yes	Yes	Yes	Yes	High
[74]	Yes	Yes	Yes	Yes	High
[75]	Yes	Yes	Yes	Yes	High
[76]	Yes	Yes	Yes	Yes	High
[77]	Partial	Yes	Partial	Yes	Medium
[78]	Yes	Yes	Yes	Yes	High
[79]	Yes	Yes	Yes	Yes	High
[80]	Yes	Yes	Yes	Yes	High
[81]	Yes	Yes	Yes	Yes	High
[82]	Yes	Yes	Yes	Yes	High
[83]	Yes	Yes	Partial	Yes	Medium
[84]	Yes	Yes	Yes	Yes	High
[85]	Yes	Yes	Yes	Yes	High

I. Data Extraction and Synthesis

For each included study, relevant technical details were systematically extracted, including the type of predictive model, forecasting horizon, data sources, model architecture, feature selection techniques, feature engineering techniques, tuning techniques, evaluation metrics, validation methods, key findings, and reported limitations. Due to variations in datasets and experimental settings, a descriptive qualitative synthesis was employed instead of a quantitative meta-analysis.

J. Classification Framework

The selected studies were organised according to their primary methodological approach into three categories: ML, DL, and Hybrid/Ensemble models. Each study was classified based on its principal contribution or best-performing model, rather than on all models included in the comparative evaluation. This methodological categorisation enables a structured comparison across studies, despite variations in objectives, datasets, and performance metrics, thereby facilitating the identification of methodological trends, research gaps, and potential directions for future work.

IV. Review of Selected Studies

In this section, 37 studies on ML, DL and Hybrid/Ensemble solar irradiance forecasting published between 2019 and 2025 are reviewed. The summaries of these studies are presented in Table III to Table V. Each research is evaluated in terms of the purpose, research methodologies, findings, and limitations found. The Goal dimension reflects the main research aim of every research, whereas the Method dimension explains the models or techniques applied. The outcome dimension will summarise the key results or findings of the performance, and the limitations dimension will identify the limitations or challenges that are reported by the authors. This is a summary of previous work that used ML, DL, and Hybrid/Ensemble in predicting solar irradiance. The methods go all the way from classical regression and tree models to DL structures, such as CNNs, RNNs, LSTMs, and hybrid/ensemble models. The works explore both the short-term (minutes to hours) and long-term predictions, using various input characteristics such as historical irradiance, meteorological variables, time indicators, sky images, and satellite-based data. Other works examine the probabilistic forecasting and quantification of uncertainty to aid in decision-making with regard to grid operations.

TABLE III. Summary of Selected ML Studies on Solar Irradiance Forecasting

Reference	Goal	Method	Outcome	Limitations
[49]	To evaluate whether linear ML algorithms can provide accurate short-term solar irradiance forecasting using sky images while maintaining low computational complexity.	Applied linear ML, including Lasso Regression, Ridge Regression (RR), Bayesian Ridge (BR), Stochastic Gradient Descent (SGD), LR, Generalized Linear Model (GLM), and RANSAC using three years (2014-2016) 1-minute resolution sky images (1536×1536 RGB) and solar irradiance data collected at the Energy Ville campus, Belgium. Used evaluation metrics including RMSE, Forecast Skill (FS), RMSE of Ramp Rate (RR), and Execution time (seconds) and used a chronological train/test split for model validation. Used Fisheye correction for feature engineering.	The Lasso achieved the highest accuracy for 5-10 min ahead forecasts, with RMSE of 0.05-0.062 W/m². SGD performed best for 15-30 min ahead with RMSE of 0.067-0.076 W/m². BR and RR consistently ranked among the top-performing models across all horizons. All models except RANSAC outperformed the smart persistence benchmark, demonstrating that simple linear models can rival more complex approaches for short-term forecasting.	Forecasting horizons are limited to 30 minutes due to the spatial constraints of sky imagers. The study is based on a single geographic location, limiting generalizability. Model performance depends strongly on image quality, reducing its applicability in regions lacking imaging infrastructure.
[50]	To develop a computationally efficient method for short-term and ahead-of-time solar irradiance forecasting from sky images while maintaining competitive accuracy.	The kNN and RF were employed for nowcasting and forecasting solar irradiance up to 4 hours ahead using two single-location sky image datasets (TSI-880 and ASI-16) covering 15 years in 10-minute resolutions, paired with GHI measurements from Golden, Colorado, USA. Used Latent Semantic Analysis (LSA) and Singular Value Decomposition (SVD) for future extraction. Used evaluation metrics, including Normalized MAPE (nMAPE) RMSE, Normalized RMSE (nRMSE) and used a chronological train/test split for model validation.	The proposed kNN model achieved competitive hourly forecasting accuracy, with nMAPE of 14.9% for TSI-880 and 14.7% for ASI-16; RMSE of 122.2 and 116.7 W/m²; and nRMSE of 7.7% and 8.9%, respectively, while reducing computational complexity by about 30% compared to state-of-the-art CNN/LSTM-based DL models.	The model's performance relies heavily on the availability and quality of sky camera images, limiting its applicability in regions lacking such infrastructure. Evaluation was conducted at a single site, limiting assessment of generalizability to other climates, camera setups, or large-scale deployment.
[51]	To forecast daily solar radiation in Riyadh, Saudi Arabia, using ML techniques and compare their predictive performance.	The study applied three ML models: ANN, RF, and LR, using historical meteorological data to predict daily GHI, obtained from the King Abdullah City for Atomic and Renewable Energy (KACARE). Used evaluation metrics including RMSE, MAE and direction accuracy and used a hold-out train/test split for model validation.	The RF model achieved the best performance, with a direction accuracy of 92.8571%, MAE of 13.6747 and an RMSE of 10.3157, outperforming the ANN and LR models. The results demonstrate that ensemble tree-based methods provide more accurate solar radiation predictions than traditional regression and neural	The study used data from a single location, limiting the generalizability of the results to other regions. Additionally, the comparison involved only three ML models, and more advanced DL approaches were not investigated. The study also focused on daily forecasting, without evaluating multi-step

			network approaches for the dataset studied.	or short-term high-resolution predictions.
[52]	To evaluate solar power forecasting performance using statistical, ML, and DL approaches.	Multiple algorithms were tested, including statistical models (Moving Average, Exponential Smoothing, ARIMA, SARIMA), ML models (RF, DT, LR, SVR, GB), and DL models (ANN, RNN, CNN, LSTM), using a dataset containing weather variables and generated solar power, for a 3-hour ahead forecasting horizon. Used Correlation-based feature selection. Feature extraction via transformations. Used evaluation metrics, including RMSE and MAE, and used a hold-out train/test split for model validation. Used adaptive feature selection	The RF model achieved the best forecasting performance with an RMSE of 4.04 kW and an MAE of 2.54 kW, outperforming all other models. DL models showed moderate performance due to the limited dataset size, and statistical models produced the highest errors, indicating poor capability to capture nonlinear patterns in solar power data.	The dataset used in the study was relatively small, limiting the performance of DL models, which typically require large datasets. The study also focused on single-dataset evaluation, meaning the results may not generalise to other geographic regions or datasets. Additionally, hybrid forecasting models were not explored, which could potentially improve prediction accuracy by combining multiple algorithms.
[53]	To improve the accuracy and reliability of solar power forecasting by evaluating different ML and DL models and identifying the most effective approach.	Four models were implemented and compared: LR, Polynomial Regression (PR), ANN, and RF, using historical solar power, meteorological, and environmental data for forecasting horizons of 1-10 days. Random Forest Regression (RFR) was used for future selection. Used evaluation metrics, including RMSE and R^2, and used a chronological train/test split for model validation. Correlation matrix used for feature selection. Grid search CV used for tuning.	The RF model achieved the best forecasting performance, with the lowest RMSE of 1.72 and the highest R^2 of 97.5% for up to 10-day-ahead forecasting. ANN also performed well but was slightly less accurate than RF. LR and PR produced higher prediction errors. The results demonstrate that ensemble learning methods, such as Random Forests, are highly effective at capturing nonlinear relationships in solar power data.	Forecasting accuracy decreases as the prediction horizon increases, especially for long-term forecasts such as 10 days ahead, due to increased weather uncertainty. The study evaluated only four models, and hybrid or ensemble combinations beyond RF were not explored. Additionally, the study does not clearly discuss dataset diversity or multiple geographic locations, which may limit the generalizability of the results.

TABLE IV. SUMMARY OF SELECTED DL STUDIES ON SOLAR IRRADIANCE FORECASTING

Reference	Goal	Method	Outcome	Limitations
[54]	To develop and evaluate DL based models for daily solar irradiance forecasting using single-site and multi-site data.	Evaluates DL models, including LSTM, GRU, CNN, Bidirectional LSTM (Bi-LSTM), and Attention-based LSTM, using multi-location and single-location solar irradiance data spanning 36 years. Pearson correlation, Spearman correlation, and XGB feature importance were used for feature selection. Used evaluation metrics including MSE, RMSE, and R^2 and used a rolling window validation method.	Demonstrated that bi-LSTM and at-LSTM models provide accurate daily solar irradiance forecasts, with R^2 values of 0.68-0.70 in the first location and R^2 values of 0.72-0.73 in the second location. This study shows that incorporating multi-site historical data improves prediction accuracy and reliability compared to single-site univariate models.	Performance gains depend on long-term, multi-site data, potentially limiting applicability in regions with sparse measurement infrastructure. Computational costs and DL model complexity were not assessed, restricting insights into efficiency, scalability, and real-world deployment.
[55]	To evaluate and compare DL and ML models for long-term solar radiation forecasting to support microgrid planning and photovoltaic energy potential assessment.	Applies multiple DL and ML models, including GRU, LSTM, RNN, FFNN, and SVR, RF on 17 years of hourly and daily solar irradiance data (2000-2017) from two regions in Korea. Used Pearson correlation analysis for feature selection. Used only RMSE evaluation metrics and used validation methods, including chronological train-test split and sliding-window method.	Demonstrated that GRU outperformed the other DL and ML models, with RMSE of between 0.3909 W/m^2 and 0.4582 W/m^2 for the hourly solar radiation data and 5.3315 W/m^2 and 6.2672 W/m^2 for the daily solar radiation data, demonstrating its effectiveness for 1-year-ahead long-term solar radiation forecasting.	The study focuses on long-term horizons, limiting applicability to short-term forecasting. Computational costs and DL model complexity versus ML models were not examined, restricting insights into efficiency, scalability, and real-world deployment.
[56]	To develop a sky image-based DL model for short-term 1-hour-ahead GHI forecasting without relying on numerical meteorological inputs across different weather conditions.	Develops a deep CNN SolarNet and compares it to the Persistence model (classic), Persistence of Cloudiness, SVR, and ANN, trained on six years of sky images and corresponding GHI data from a single site. Used nRMSE and FS evaluation metrics, and used a chronological train-test split validation method	The SolarNet model demonstrated robust performance, outperforming the other models, achieving an nRMSE of 8.85% and a forecasting skill score of 25.14%.	Evaluation was limited to a single site, raising concerns about generalizability. The approach relies solely on image-based inputs, excluding numerical meteorological variables. The high computational complexity of the deep CNN architecture may limit its applicability in regions without sky-image infrastructure.

[57]	To compare the predictive performance of ML and DL models for Global Solar Radiation (GSR) and Diffuse Solar Radiation (DSR) prediction across multiple locations.	Evaluates and compares several ML and DL models, including SVR, RF, Polynomial Regression (PR), ANN, CNN, and RNN, using 12 years of hourly solar irradiance data (2005-2016) from four locations in Nigeria. Used evaluation metrics including *r*, RMSE, MAE, NMBE and used a chronological train-test split validation method	The DL models outperformed the ML models for GSR and DSR prediction, with the RNN achieving the best overall performance in Yobe, yielding *r* of 0.9546, RMSE of 82.22 W/m², and MAE of 36.52 W/m². ML models had the lowest computation costs.	Computational costs and DL model complexity relative to ML models were not examined, limiting understanding of their efficiency, scalability, and real-world applicability.
[58]	To compare statistical, ML, and DL models for short-term and long-term forecasting of GHI using high-quality measured data from Islamabad, Pakistan.	Compares multiple forecasting models, including Seasonal ARIMAX, Facebook Prophet (FBP) for longer horizons, and ML and DL models such as ANN, CNN, and LSTM on 4.75 years of hourly GHI data (2017-2021) from Islamabad, Pakistan. Used evaluation metrics including R^2, MAE, MSE, RMSE and used a time-ordered train-test split validation method. Used Pearson correlation for feature selection.	The DL models performed best for short-term GHI forecasts, achieving R2 scores of 0.987, 0.977, and 0.984, while statistical models were more efficient for long-term forecasts, with R^2 scores above 0.90.	The study is limited to a single geographic location, restricting generalizability to other climates. Computational costs and DL model complexity relative to ML models were not assessed, limiting insights into efficiency, scalability, and real-world deployment.
[59]	To develop a multi-modal fusion network for short-term solar irradiance micro-forecasting using both infrared images and past irradiance data, addressing intermittent and nonlinear patterns during cloudy periods.	Proposes a CNN-L and Multiple-Image CNN-LSTM fusion network (MICNN-L) that integrates infrared sky/cloud images and historical irradiance time-series and is evaluated against benchmark models, including Linear SVM, Nonlinear SVM, Linear GP, Nonlinear GP, DNN, RNN, GRU, and LSTM using 15-second infrared sky images and solar irradiance data from the University of New Mexico, Albuquerque, USA. Used evaluation metrics, including MAPE, and used a time-ordered train-test split validation method.	MICNN-L showed 46.42% improvement in MAPE, and CNN-L showed 42.02% improvement in MAPE compared to benchmark ML/DL models. Cloudy day classification achieved 99.23% accuracy. Proposed models effectively captured nonlinear spatio-temporal variations and improved short-term forecasts.	Evaluation is based on single-site, high-frequency (15 s) data, with dataset duration and calendar years unspecified. Reliance on infrared sky imagery increases computational complexity and limits applicability in regions without sky-image infrastructure.
[60]	To compare ML and DL algorithms for solar irradiance forecasting and identify the most effective model.	Evaluated several ML models, including DT, kNN, RF, SVM, XGB, Stacking, Light GBM, Boosting, CatBoost, MLP, and LSTM, on three years (2017-2020) of hourly and 15-minute solar irradiance data for multiple locations in Izmir, Turkey. Used evaluation metrics, including MAE, MSE, RMSE, and R^2 and used a chronological train-test split validation method. Tested LightGBM, Pearson, recursive elimination, chi-square, and LR for feature selection.	The MLP achieved the best performance with an RMSE of 0.79, demonstrating that neural networks can effectively capture nonlinear patterns in GHI forecasts and guiding solar panel deployment decisions.	The dataset is specific to the NSRDB and spatially restricted to a 4 km × 4 km area, limiting generalizability. In addition, computational cost and relative model complexity were not assessed, constraining insights into efficiency, scalability, and real-world deployment.
[61]	To accurately forecast solar electricity generation at a solar farm in Amherst for operational, technological, and financial planning.	Develops and compares several DL variants, including LSTM-V, Bi-LSTM, and GRU-II, as well as baseline regressors such as LASSO, Ridge, Elastic-Net, and SVR on four-year (2015-2019) 5-minute interval time-series meteorological and solar energy data from the Amherst weather station in Massachusetts, USA. Used evaluation metrics, including RMSE, MSE, MAE, and MAPE and used a hold-out train-test split validation method. Used a correlation heatmap for feature selection	The Bi-LSTM achieved the best performance, outperforming the other models, with an MAE of 0.0135, RMSE of 0.0315, MSE of 0.0012, and MAPE of 0.1205.	The dataset is restricted to a single location, limiting generalizability across different climatic conditions. Only five-minute interval predictions were evaluated; longer forecasting horizons were not assessed. Additionally, computational cost and relative model complexity were not examined, constraining insights into efficiency, scalability, and real-world deployment.
[62]	To accurately forecast short-term solar irradiance from historical sky image sequences while leveraging clear sky models for improved performance.	Employed a DL Vision Transformer-based network with auxiliary clear sky model input, benchmarked on two datasets spanning from 2004 to 2019 in 10-minute intervals. Used the evaluation metric normalised Mean Absolute Percentage (nMAP) and used a chronological train-test split validation method.	The DL Vision Transformer outperforms prior DL methods across two datasets, achieving reduced nMAP errors by 6.9% and 8% for nowcasting for the 2015 and 2016 datasets, and 5.4% and 8.8% reduction for 4-hour ahead prediction, and outperformed previous methods, especially for 3-	The study is limited to benchmark datasets (TSI880 and ASI16), with performance under diverse climatic conditions or alternative camera configurations not evaluated. The forecasting horizon is limited to 4 hours, and computational costs were

			and 4-hour forecasts under variable weather conditions.	not assessed, limiting insights into scalability and practical deployment.
[63]	To evaluate the impact of different forecasting time horizons on the accuracy of solar irradiance prediction using ML and DL models.	Compares eight models, including LR, CNN-LSTM, XGB, RF, CNN, BiLSTM, ConvLSTM, and MLP, on 2 years (2020-2021) of solar irradiance data at 60 min, 30 min, 15 min, 10 min, and 5 min intervals in Limpopo Province, South Africa. Used evaluation metrics, including RMSE, nRMSE, MAE, and R^2 and used a hold-out train-test split validation method.	The results showed that prediction accuracy generally decreases with increasing prediction horizon and that DL models outperformed conventional ML models overall. The ConvLSTM model achieved the best RMSE of 7.43 at a five-minute interval, and the MLP performed better than other models across most prediction intervals, with RMSE ranging from 18.87 to 75.52.	The dataset is restricted to a single location, limiting generalizability across different climatic conditions. Model performance for longer forecasting horizons (>1 hour) was not evaluated. Additionally, computational cost and relative model complexity were not assessed, constraining insights into efficiency, scalability, and real-world deployment.
[64]	To explore DL techniques for solar energy forecasting within microgrid systems.	Evaluates multiple ML and DL models for forecasting solar irradiance, including Baseline Regression, LR, Lasso, Ridge, Elastic-Net, RF, GB, XGB, LightGBM, RNN, GRU, and LSTM on 9.11 years of hourly solar irradiance data from a chosen location. Used the evaluation metric normalised RMSE and used a hold-out train-test split validation method.	The DL RNN outperformed the other models, with a testing RMSE of 0.00534, and the Traditional ML RF model also performed well, with a testing RMSE of 0.09.	Computational cost and relative model complexity were not evaluated, limiting insights into practical efficiency, scalability, and deployment feasibility. Additionally, the dataset is restricted to a single location, constraining generalizability across different climatic conditions.
[65]	To evaluate DL models for short-term 1-hour ahead solar irradiance forecasting.	Compares several DL models such as Temporal Convolutional Networks (TCN), ARIMAX, SVR, standard LSTM, LSTM with 32 convolution kernels, and SENet using 3 years (2013-2016) of hourly solar irradiance data from Bend, Oregon, USA (University of Oregon). Used the evaluation metric normalised RMSE, Training time per epoch, and GPU memory usage. Used a cross-year validation method.	LSTM achieved the highest forecasting accuracy, with an average RMSE of 0.069, while TCN significantly reduced training time and resource consumption, increasing only 6% to an average RMSE of 0.0738.	Evaluation was restricted to 1-hour-ahead forecasts, with performance for longer horizons or alternative datasets not assessed. Additionally, the dataset is limited to a single location, constraining generalizability across different climatic conditions.
[66]	To compare DL models for very short-term solar irradiance forecasting and identify the most accurate architecture for local PV applications.	Evaluates RNN, GRU, LSTM, and TCN architectures for one-step-ahead forecasting using 1-year (2019) 15-minute resolution solar irradiance data from Karachi, Pakistan, obtained from NSRDB. Used evaluation metrics, including RMSE, MAE, and R^2 and used a hold-out train-test split validation method.	The LSTM model outperformed other models, achieving the highest R^2 of 0.9934 and the lowest RMSE of 0.020051, demonstrating its effectiveness for highly accurate solar irradiance forecasting.	The dataset is limited to a single location, restricting generalizability across varying climatic conditions. Computational cost and model complexity were not assessed, limiting insights into efficiency, scalability, and real-world deployment. In addition, performance for longer forecasting horizons (>15 minutes) was not evaluated.
[67]	To develop a DL based approach for short-term solar irradiance forecasting using sky camera images.	The study indeed proposes an LSTM-based DL model compared against the Yoo model and the persistence model using 25 days of 1-minute-resolution sky images from a ground-based sky camera, combined with GHI measurements at Kookmin University in Seoul, South Korea. Used evaluation metrics, including MBD, RMSD, rMBD, rRMSD and UVI and used a hold-out train-test split validation method.	The proposed LSTM-based DL model achieves better 10-minute-ahead forecasting accuracy under partly cloudy conditions (UVI > 20), with RMSD of 199.75 W/m², rRMSD of 25.1%, MBD of -60.65 W/m², and rMBD of -12.70%, outperforming the persistence and Yoo models. For clear and overcast sky conditions (UVI < 5), the model's performance is comparable to that of the persistence model, reflecting the lower irradiance variability and simpler forecasting requirements in these cases.	The model's performance relies heavily on the availability and quality of sky camera images, limiting its applicability in regions lacking such infrastructure. Evaluation was conducted at a single site, limiting assessment of generalizability to other climates, camera setups, or large-scale deployment.
[68]	To investigate whether DL models can improve the accuracy of solar radiation forecasting for advanced building	Developed and compared an ANN model and an RNN model for solar radiation prediction, using a one-week dataset with different sampling frequencies (1-hour	The RNN significantly outperformed the ANN, achieving 47% improvement in NMBE and 26% in RMSE. Increasing data granularity (from 1 h to 10 min)	The study relied on a relatively short historical dataset, limiting the model's ability to capture long-term patterns and seasonal variability. The study did not

	control applications, such as model predictive control (MPC).	and 10-minute data) from a local weather station in Alabama for training. Used evaluation metrics, including RMSE CV(RMSE), and nMBE and used a hold-out train-test split validation method. Used Pearson correlation for feature selection.	reduced CV(RMSE) from 30.9% to 9.41% for ANN and from 9.83% to 7.64% for RNN. Implementing a moving-window algorithm further improved RNN nMBE from 0.9% to 0.2%. Cloud cover was identified as a significant influencing factor on prediction accuracy.	provide a quantitative evaluation of computational complexity, training time, or scalability, making it unclear how practical the models are for large-scale or real-time deployment. Validation was conducted under offline experimental settings without real-time analysis, so model performance under operational conditions remains untested.
[69]	To develop and evaluate a near-real-time solar irradiance forecasting model using only historical GSR time series in a data-scarce region of Vietnam.	Developed a standalone DL optimised LSTM network compared to ARIMA, DNN, MLP, and SVR, and evaluated across multiple short-term forecast horizons (1-30 min) using 1-min resolution GSR time series from Bac-Ninh, Vietnam. Used evaluation metrics, including *r*, RMSE, ENS, MAE, RRMSE, and MAPE and used a hold-out train-test split validation method. Used PACF-based lag feature selection. Used grid search for tuning.	The LSTM model consistently outperformed benchmark models across all horizons. At a 1-minute horizon, it achieved RMSE of 40.91 W/m², MAE of 21.64 W/m², R of 0.9920, WI of 0.9984, and ENS of 0.9831. Relative RMSE ranged from 10 to 15% across horizons.	The dataset is restricted to a single location, limiting generalizability across different climatic conditions. Computational cost and model complexity were not evaluated, constraining insights into efficiency, scalability, and practical deployment. Validation restricted to very short-term horizons
[70]	To improve solar irradiation forecasting accuracy using ML and DL models, and evaluate the effectiveness of the N-BEATS architecture for time-series solar radiation prediction.	Evaluated SARIMA, LSTM, and N-BEATS (Neural Basis Expansion Analysis for Time Series Forecasting) to predict clear-sky GHI, DHI, and DNI using daily aggregated meteorological data from the NSRDB for four regions in Bangladesh (Khulna, Chittagong, Sylhet, and Rajshahi). Used evaluation metrics, including MAE, RMSE, R^2, and MAPE and used a hold-out train-test split validation method. Used XGB feature importance for feature selection.	The N-BEATS model achieved the highest forecasting accuracy, with RMSE values of 29.03-35.77 for clear-sky GHI and MAPE values of 2.34-3.29%, outperforming SARIMA and LSTM. N-BEATS effectively captured solar radiation patterns, while LSTM performed the worst.	The study was limited to four regions within Bangladesh, which restricts the generalisability of the findings to other climatic conditions. Additionally, the dataset was aggregated to daily values, potentially overlooking short-term variability in higher-resolution solar radiation data. The study also evaluated a limited number of forecasting models, and hybrid or ensemble approaches were not explored.
[71]	To conduct a comparative analysis of ML and DL strategies for solar irradiation forecasting.	The study evaluated 9 ML and DL models, including LR, MLP, ELM, Kernel Ridge Regression (KRR), SVR, Gaussian Process Regression (GPR), CNN, LSTM, and Bi-directional LSTM (Bi-LSTM), using historical solar irradiation data from 2000-2005, obtained from the Batna meteorological station in Algeria. Used evaluation metrics, including MAE, RMSE, and R, and used a chronological train-test split validation method.	Among ML models, ELM with 16 features achieved the best performance, with RMSE of 4.004, MAE of 2.918, and R of 0.849. Among DL models, Bi-LSTM achieved the best overall performance for 1-day-ahead forecasts, with RMSE of 1.392, MAE of 1.021, and R of 0.989 using 25 features. The results demonstrate that DL models, particularly Bi-LSTM, significantly outperform traditional ML methods for solar irradiation forecasting.	The study used data from only one location, which limits the generalisability of the results to other climatic regions. Additionally, the study focused on individual ML and DL models, rather than hybrid or ensemble approaches, which could further improve forecasting accuracy.

TABLE V. SUMMARY OF SELECTED HYBRID/ENSEMBLE STUDIES ON SOLAR IRRADIANCE FORECASTING

Reference	Goal	Method	Outcome	Limitations
[72]	To compare the performance of common ML and DL algorithms for short-term solar irradiance forecasting and improve prediction accuracy under irradiance variability.	Evaluates ML models, including SVR, kNN, RF, and DL models such as ANN, CNN, LSTM, GRU, and a hybrid CNN-LSTM model, using grid-search hyperparameter optimisation on five years of hourly solar irradiance data (2015-2019) from a single site in Islamabad. Used evaluation metrics including adjusted R^2, NRMSE, MAD, MAE, and MSE. Used a 5-fold grid search cross-validation method for tuning.	Demonstrated that the CNN-LSTM hybrid model achieved the best performance among the other models with an adjusted R^2 of 0.984.	The study is based on a single geographic location and a fixed historical dataset, raising concerns about its generalizability to other regions.
[73]	To develop a robust and accurate ensemble model for hourly GHI forecasting to support reliable solar energy	Develops an XGBF-DNN hybrid ensemble model and compares it to Smart Persistence (SP) and other ML models, including SVR, RF, XGB, and DNN, on 10 years of hourly solar irradiance data (2005-2014) across three different climatic regions in India. Used evaluation metrics including FS,	Demonstrated that the proposed XGBF-DNN model outperformed the other models for a 1-hour ahead forecast, with RMSE values	The study focuses on long-term horizons, limiting applicability to short-term forecasting. Computational costs and DL model complexity relative to ML

	system planning and operation.	RMSE, and MAE. Used a Clear-sky index and temporal encoding feature engineering.	of 56.68 W/m^2, 53.78 W/m^2 and 91.86 W/m^2 for all three locations, with 33-40% improvement in FS.	models were not assessed, restricting insights into efficiency, scalability, and real-world deployment.
[74]	To improve 1-hour-ahead solar irradiance forecasting accuracy by incorporating multivariate inputs and a multi-branch hybrid DL structure.	Develops a hybrid WPD-CNN-LSTM-MLP model using a multi-branch architecture and compares it to BP, RNN, LSTM, SVM, BP-MLP, RNN-MLP, LSTM-MLP, and CNN-LSTM-MLP on 3 years of hourly solar irradiance data (2012-2016) across three sites in the USA: Denver, Clark, and Folsom. Used evaluation metrics including RMSE, MAE, MBE, FS and R^2. Used WPD decomposition feature engineering.	Demonstrated that the proposed WPD-CNN-LSTM-MLP model achieved superior prediction accuracy compared to other models, with MAE of 14.3523-20.3853, RMSE of 32.1006-46.1336, nRMSE of 15.4795-23.9178%, and R values of 0.9858-0.9940 across the three locations. WPD further reduced errors and improved correlation.	The study focuses on short-term horizons, limiting applicability to long-term forecasting. Computational costs and DL model complexity relative to ML models were not assessed, restricting insights into efficiency, scalability, and real-world deployment.
[75]	To develop an accurate and robust short-term hourly GHI forecasting model for reliable grid integration and energy market participation.	Develops a hybrid LSTM-CNN DL model and compares it to SP, SVM, ANN, CNN, LSTM, and CNN-LSTM on 1 year of hourly solar irradiance data (2019) across 23 locations. Used evaluation metrics including RMSE, MAE, R^2 and computational time.	The hybrid LSTM-CNN model improved prediction and outperformed standalone models for 1-hour-ahead forecasts, with MAE of 27.38-37.02, RMSE of 42.89-58.12, R of 0.9673-0.9853, and FS of 37%-45% across all the locations.	The study focuses on short-term horizons, limiting applicability to long-term forecasting. Computational costs of the hybrid DL model relative to ML models had high computational times, restricting efficiency, scalability, and deployment in computational resource-constrained locations.
[76]	To develop a hybrid short-term solar irradiance forecasting model for both GHI and Plane of Array (POA) irradiance with high predictive accuracy.	Develops a modified multi-step CNN-stacked LSTM network with dropout and compares it to CNN and LSTM on one month of solar irradiance data (2019) at a single site from Sweihan Photovoltaic Independent Power Project, Abu Dhabi, UAE. Used evaluation metrics including RMSE, MAE, MAPE, R^2, MSE, and NRMSE. Used a hold-out train-test split validation method.	The proposed hybrid model outperformed other ML and DL models for 1-hour-ahead forecasts, achieving an RMSE of 0.36, an R^2 of 0.98 for GHI, an MAPE of 61.24, and an R^2 of 0.96 for POA irradiance, demonstrating improved accuracy for short-term solar irradiance prediction.	The study is limited to a single location, affecting generalizability, and the computational costs and complexity of DL models versus ML models were not assessed, limiting insights into efficiency, scalability, and real-world deployment.
[77]	To develop a hybrid intelligent model for accurate short-term solar irradiance/power forecasting using wavelet decomposition and Generative Adversarial Networks (GANs) with evolutionary optimisation for two regions.	Proposes a Wavelet Transform Package (WTP)-GAN, optimised using a modified Dragonfly Algorithm (DA), and compared to ARMA, ANN, CNN, GAN, SVR, Fuzzy method, and RNN using 1 year of hourly solar irradiance data (2019) from two locations. Used evaluation metrics including RMSE and MAPE. Used a Wavelet decomposition feature engineering. Used dragonfly algorithm plus modified 3-phase DA for hyperparameter tuning.	The hybrid WTP-GAN with modified DA outperformed other models across multiple horizons of 1-6 hours ahead, with MAPE of 0.0282 & 0.0262, and RMSE of 0.0473 & 0.0479 for 1-pace forecast, and MAPE of 0.0531 & 0.0631, and RMSE of 0.0895 & 0.0946 for 6-pace forecast:	The study relies on a single year of data, limiting seasonal and multi-year generalizability. Moreover, computational cost and model complexity were not assessed, restricting insights into efficiency, scalability, and practical deployment.
[78]	To develop a heterogeneous ensemble dynamic selection (HetDS) model for solar irradiance forecasting, improving performance over single models.	Proposes a HetDS model that chooses the best forecasting model for each test pattern from a pool of seven methods, which include ARIMA, SVR, ANN, Extreme Learning Machine (ELM), Deep Belief Network (DBN), RF, and GB, using 1 year (2020) hourly solar irradiance data from four locations in Brazil. Used evaluation metrics including RMSE, MAPE, and MAE. Used a time-lag windowing and feature engineering. Used full Grid Search for hyperparameter tuning. Used a chronological train/test split validation.	The HetDS models performed better than the other models for 1-hour-ahead forecasts, with RMSE ranging from 0.0600 to 0.1762, MAPE ranging from 10.57% to 55.85%, and MAE ranging from 0.0426 to 0.1169 across all four cities.	The study is based on one year of data, limiting seasonal and multi-year generalizability. Additionally, computational cost and relative model complexity were not evaluated, constraining insights into efficiency, scalability, and real-world deployment.
[79]	To develop a hybrid decomposition-integration DL framework for short-term solar irradiance forecasting that better captures nonlinear and non-stationary patterns in irradiance data.	Develops a Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)-Wasserstein GAN-LSTM hybrid model and compares to benchmark models evaluated, including RNN, GRU, Transformer, LSTM, WGAN, CEEMDAN-LSTM, CEEMDAN-WGAN, CEEMDAN-LSTM-WGAN variants using one year (2018), 5-minute resolution	The proposed CEEMDAN -WGAN-LSTM model outperformed all benchmark models across four seasonal test sets. Compared to sub-optimal benchmarks, its errors were reduced by 3.51% for MAE, 6.11% for MAPE, and 2.25%	The model was evaluated using one year of data from a single location, limiting generalizability across different climatic regions and multi-site settings. In addition, computational cost and relative model complexity were not assessed, constraining

		GHI data from NSRDB measured in Texas, USA. Used evaluation metrics including MAE, MAPE, RMSE, and R^2. Used a CEEMDAN-based decomposition for feature engineering and feature selection. Used a chronological train/test split validation.	for RMSE, and it showed the lowest errors with the highest R^2 values among the evaluated methods across seasons.	insights into efficiency, scalability, and real-world deployment.
[80]	To improve mean hourly GHI forecasting by providing accurate probabilistic predictions for grid stability and PV integration.	Develops a hybrid CNN-LSTM model combined with quantile regression and compares it with ANN, CNN, LSTM, and persistence models using one year (2019-2020) of hourly solar irradiance data, including sky images down-sampled to 64×64 pixels collected at the University of French Polynesia, Tahiti. Used evaluation metrics including MAE, RMSE, MSE, R^2, nMAP (deterministic); PICP, PINAW, CWC (probabilistic). Used a Sky image down-sampling for feature engineering. Used a chronological train/test split validation.	The CNN-LSTM hybrid model achieved the best deterministic and probabilistic performance for 1-h-ahead forecasts, with an RMSE of 100.58 W/m², an MAE of 66.09 W/m² and an R^2 of 0.85 on test data. For probabilistic forecasting, it obtained the lowest CWC values across multiple prediction intervals: 8.54 at PI (38%), 17.04 at PI (68%), 33.37 at PI (95%), 47.98 at PI (99%), demonstrating superior uncertainty quantification.	The dataset is restricted to a single location, limiting generalizability across different climatic conditions. Computational cost and model complexity were not evaluated, constraining insights into efficiency, scalability, and practical deployment. In addition, reliance on satellite imagery reduces applicability in regions without imaging infrastructure.
[81]	To improve 24-hour-ahead solar GHI forecasting accuracy for reliable grid management using an ensemble DL approach.	Employed a hybrid Wavelet Transform (WT)-BiLSTM model and compared it with GRU, LSTM, and standalone BiLSTM models using a one-year monthly solar GHI dataset, using a one-month moving window mechanism for the location of Ahmedabad, Gujarat, India. Used evaluation metrics including MAPE, RMSE, FS, and R^2. Used a moving window mechanism validation.	The proposed WT-BiLSTM ensemble outperformed benchmark, LSTM, GRU, and standalone BiLSTM models for 1-day-ahead forecasts, with a reduced Monthly average RMSE of 26.04-58.89% vs benchmark and up to 57% vs LSTM. The MAPE reduced by 9-51.18% vs benchmark. The achieved R^2 is 0.94, and the forecast skill is 47%.	The dataset is restricted to a single location, limiting generalizability across different climatic conditions. Computational cost and model complexity were not evaluated, constraining insights into efficiency, scalability, and practical deployment.
[82]	To improve solar radiation forecasting accuracy in Indian cities by integrating advanced signal decomposition techniques with ML and DL models.	This study developed nine novel ensemble models integrating Variational Mode Decomposition (VMD) and Discrete Wavelet Transform (DWT) with five DL algorithms, including GRU, LSTM, BiLSTM, CNN, and DNN, along with two ML algorithms, ANN and SVR, using historical meteorological data from 2000 to 2012 in seven major Indian cities provided by the Indian Meteorological Department. Used evaluation metrics including RMSE, MAE, and R^2. Used a VMD & DWT signal decomposition for feature engineering.	VMD-based DL ensembles significantly outperformed DWT-based models. Among all configurations, the VMD-GRU model achieved the best performance across all cities for 1-step-ahead forecasts, with RMSE of 0.82-1.22, MAE of 0.54-1.02, and R^2 of 0.83-0.93. This superior performance is attributed to GRU's lower parameter count, reduced memory requirements, and faster training capability.	The study utilised a limited set of meteorological variables, excluding other potentially influential atmospheric parameters. The study did not provide quantitative evaluations of computational complexity, training time, or the scalability of the hybrid models. Validation was conducted in an offline experimental setting without real-time deployment analysis.
[83]	To enhance solar radiation forecasting accuracy by developing a hybrid intelligent DL framework that integrates signal decomposition and evolutionary optimisation techniques.	Proposes a hybrid OVMD-PACF-ISSA-DBN-OSELM hybrid model combining optimal VMD with an evolutionary DBN online sequential ELM (EDBN-OS-ELM); applies partial autocorrelation function (PACF) for input variable selection; improves the sparrow search algorithm (SSA) to optimise DL hyperparameters. The model was developed and validated using hourly solar radiation data from January, May, and October 2022, collected from a single station of the National Data Buoy Centre. Used evaluation metrics including RMSE, MAE and R. Used hold-out train/test split validation method. Used the Partial Autocorrelation Function for feature selection. Used OVMD (Optimal Variational Mode Decomposition) for feature engineering. Used the improved Sparrow Search Algorithm (ISSA) algorithm for hyperparameter tuning.	The proposed model achieved the best performance across the January, May, and October datasets. For January, the model obtained an RMSE of 23.153 W/m², an MAE of 13.225 W/m², and an R of 0.9923, while the model obtained an RMSE of 39.223 W/m², an MAE of 24.962 W/m², and an R of 0.9924 for May. The model significantly outperformed BP, ELM, LSTM, DBN-based variants, and SSA-based hybrids for 10-minute-ahead forecasts.	The experimental evaluation was conducted using data from only three months (January, May, and October 2022), which limits the assessment of seasonal robustness. Additionally, the dataset was obtained from a single NDBC station, restricting geographical generalizability. The study focused solely on single-step forecasting, and multi-step prediction scenarios were not investigated.
[84]	To develop a robust and accurate deep hybrid	Developed a hybrid CEEMDAN-AG-RE-ELM that combines CEEMDAN for signal decomposition,	The proposed model achieved strong forecasting accuracy	The framework involves multiple complex stages

	framework for short-term GSR forecasting by combining signal decomposition, time-frequency analysis, deep feature extraction, and ML.	Continuous Wavelet Transform (CWT) to generate time-frequency scalogram images; cascade CNN (AlexNet + GoogLeNet) for deep feature extraction; RReliefF for feature selection; ELM for final forecasting using a two-year (2018-2019), hourly dataset from the Yulara Solar System, near Uluru (Ayers Rock). Used evaluation metrics including RMSE, MAE, and MAPE. Used a hold-out train/test split validation method. Used the RReliefF for feature selection. Used CEEMDAN decomposition for feature engineering. Used the 10-fold CV for hyperparameter tuning.	across multiple horizons of 1-, 2- and 3-hour ahead, with RMSE of 0.0642, MAE of 0.0241, MAPE of 0.1201 for 1-hour ahead and RMSE of 0.0686, MAE of 0.0285, MAPE of 0.1279 for 2-hour ahead and RMSE of 0.0724, MAE of 0.0315, and MAPE of 0.1317 for 3-hour ahead, outperforming conventional regression models; demonstrates superior robustness and accuracy.	(decomposition, image transformation, CNN feature extraction), which may increase computational cost. Performance was validated on a specific dataset; generalizability to other locations and longer horizons was not assessed.
[85]	To improve daily GSR prediction accuracy by developing a hybrid DL-ML framework that effectively extracts temporal features and enhances regression performance.	Proposed a hybrid CNN-SVR (CSVR) model and compared it to STM, DBN, RBF, BRF, MARS, WKNNR, GPML, and M5TREE models using Ground-measured daily GSR. Used evaluation metrics including RMSE, MAE, Willmott's d, NSE, and Legates-McCabe Index. Used a hold-out train/test split validation. Used Atom Search Optimisation (ASO) for feature selection. Used HyperOpt (Bayesian/TPE for hyperparameter tuning.	The proposed CSVR model consistently outperformed benchmark DL and ML models across six solar farms in Queensland, Australia. It achieved RMSE/MAE ranges of 2.172-3.305 MJ/m² and 1.624-2.370 MJ/m², respectively, compared to 2.514-3.879 MJ/m² and 1.939-2.866 MJ/m² from competing models. Higher correlation, Willmott's Index, Nash-Sutcliffe efficiency, and Legates-McCabe indices confirmed superior agreement between predicted and observed GSR.	Model complexity is relatively high due to multi-stage optimisation (feature selection, CNN, SVR, hyperparameter tuning). Evaluation was limited to daily forecasting and a specific geographical region, which may affect generalizability to other climates or shorter forecasting horizons.

Table VI summarises the quantitative findings of all 37 studies reviewed, indicating per reference, the model type, the best-performing model, forecast horizon, the type of input data used, the evaluation measure, and the best-reported score. This overview allows the direct cross-study comparison of the reported performance.

TABLE VI. SUMMARY OF SELECTED HYBRID/ENSEMBLE STUDIES ON SOLAR IRRADIANCE FORECASTING

Ref	Model Type	Best Model	Forecast Horizon	Input Dataset Type	Best Metric	Best Score
[49]	ML	Lasso Regression	5-30 min (Very short-term)	Numerical (irradiance + meteorological)	RMSE	0.050-0.076 W/m²
[50]	ML	kNN	4 h (Short-term)	Numerical (time series)	nRMSE	7.7-8.9%
[51]	ML	RF	1 day (Short-term)	Numerical (irradiance)	Directional Acc.	92.86%
[52]	ML	RF	3 h (Short-term)	Numerical	RMSE	4.04 kW
[53]	ML	RF	1-10 days (Medium-term)	Numerical	RMSE	1.72
[54]	DL	Bi-LSTM / Attention LSTM	1 day (Short-term)	Numerical (long-term time series)	R^2	0.68-0.73
[55]	DL	GRU	1 year (Long-term)	Numerical (hourly/daily)	RMSE	0.3909-0.4582
[56]	DL	SolarNet (CNN)	1 h (Very short-term)	Sky images	nRMSE	8.85%
[57]	DL	RNN	1 h (Short-term)	Numerical (multi-site)	RMSE	82.22 W/m²
[58]	DL	LSTM/CNN	1 h (Short-term)	Numerical (irradiance + meteorological)	R^2	0.987
[59]	DL	MICNN-L	15 sec (Very short-term)	Sky images	MAPE improvement	46.42%
[60]	DL	MLP	15 min-1 h (Very short-term)	Numerical (NSRDB)	RMSE	0.79
[61]	DL	Bi-LSTM	5 min (Very short-term)	Numerical	MAE	0.0135
[62]	DL	Vision Transformer	Now-4 h (Short-term)	Sky images + clear-sky model	nMAP	↓5.4-8.8%
[63]	DL	ConvLSTM	5-60 min (Very short-term)	Numerical	RMSE	7.43
[64]	DL	RNN	1 h (Short-term)	Numerical	nRMSE	0.00534

[65]	DL	LSTM	1 h (Short-term)	Numerical	nRMSE	0.069
[66]	DL	LSTM	15 min (Very short-term)	Numerical (NSRDB)	R^2	0.9934
[67]	DL	LSTM	10 min (Very short-term)	Numerical	RMSD	199.75 W/m²
[68]	DL	RNN	10 min-1 h (Very short-term)	Numerical	RMSE improvement	26%
[69]	DL	LSTM	1-30 min (Very short-term)	Numerical	RMSE	40.91 W/m²
[70]	DL	N-BEATS	1 day (Short-term)	Numerical (multi-region)	RMSE	29.03-35.77
[71]	DL	Bi-LSTM	1 day (Short-term)	Numerical (irradiance)	RMSE	1.392
[72]	Hybrid	CNN-LSTM	1 h (Short-term)	Numerical	Adjusted R^2	0.984
[73]	Hybrid	XGBF-DNN	1 h (Short-term)	Numerical (multi-region)	RMSE	53.78-91.86
[74]	Hybrid	WPD-CNN-LSTM-MLP	1 h (Short-term)	Numerical	RMSE	32.10-46.13
[75]	Hybrid	LSTM-CNN	1 h (Short-term)	Numerical (multi-site)	MAE	27.38-37.02
[76]	Hybrid	CNN-stacked LSTM	1 h (Short-term)	Numerical	RMSE	0.36
[77]	Hybrid	WTP-GAN + DA	1 h (Short-term)	Numerical	MAPE	0.0262-0.0282
[78]	Hybrid	HetDS	1 h (Short-term)	Numerical (multi-city)	RMSE	0.0600-0.1762
[79]	Hybrid	CEEMDAN-WGAN-LSTM	5 min (Very short-term)	Numerical (NSRDB)	MAE improvement	3.51%
[80]	Hybrid	CNN-LSTM + Quantile Reg.	1 h (Short-term)	Sky images + numerical	RMSE	100.58 W/m²
[81]	Hybrid	WT-BiLSTM	24 h (Short-term)	Numerical	RMSE reduction	26-59%
[82]	Hybrid	VMD-GRU	1 h (Short-term)	Numerical (multi-city)	RMSE	0.82-1.22
[83]	Hybrid	OVMD-ISSA-DBN-OSELM	1 h (Short-term)	Numerical	RMSE	23.15-39.22
[84]	Hybrid	CEEMDAN-AG-RE-ELM	1-3 h (Short-term)	Numerical + time-frequency	RMSE	0.0642-0.0724
[85]	Hybrid	CNN-SVR	1 day (Short-term)	Numerical	RMSE	2.17-3.31

V. Discussion

The review summarises 37 peer-reviewed publications published between 2019 and 2025, narratively summarised in Table III to Table V and quantitatively summarised in Table VI. Taken together, these studies provide a comprehensive overview of the state of the art in ML, DL and hybrid forecasting of solar irradiance. Despite the literature showing significant gains in predictive accuracy and methodological sophistication, a critical cross-analysis of the synthesised evidence reveals several key observations about current trends in methodological approach, limiting factors and important research gaps. The gaps need to be addressed to ensure operational preparedness and the generalizability of forecasting models for integrating renewable energy in the real world.

A. Key Findings and Empirical Observations

The initial important observation is that DL and hybrid models have evidently dominated the methodological approach to solar irradiance forecasting. A summary of the information in Table III, Table IV, and Table V, as shown in Fig. 2, indicates a significant shift towards complex neural network-based techniques. Fig. 2 reveals that 49% (n=18) of the studies employ DL models as reported in [54]-[71], 38% (n=14) use hybrid or ensemble approaches as reported in [72]-[85], and 14% (n=5) rely on traditional ML models as reported in [49]-[53]. Overall, this highlights a clear shift in the literature towards more data-intensive, architecture-rich forecasting approaches.

The second important observation concerns the relative performance of ML, DL, and hybrid models across various forecasting horizons and evaluation metrics. An overview of the findings in Table VI, which provides a cross-study comparative summary, shows that DL and hybrid models tend to be superior to traditional ML methods, especially for very short- and short-term forecasting. Table VI shows that the LSTM, CNN, and their variants are the best DL-based predictors across various datasets. For example, the RMSE values were as low as 7.43 [63] and 29.03-35.77 [70] when using DL models, while the coefficients of determination were as high as $R^2 = 0.987$ in [58] and $R^2 = 0.9934$ in [66]. On the same note, MAE as low as 0.0135 is reported in [61], indicating a high predictive accuracy for very short-term forecasts. Conversely, traditional ML models exhibit relatively low consistency in performance, with RMSE values of 4.04 kW [52] and 1.72 kW [53] for medium-term forecasting and nRMSE values between 7.7% and 8.9% [50]. Hybrid and ensemble models also improve predictive performance by integrating various learning paradigms or signal decomposition methods. As shown in Table VI, hybrid models frequently achieve competitive or superior results, with RMSE values of 0.36 [76], 0.0600-0.1762 [78], and 23.15-39.22 [83], as well as MAE values of 27.38-37.02 [75]. Moreover, other hybrid methods do not provide absolute values but enhanced performance, such as RMSE decreases of 2659% [81] and MAE increases of 3.51% [79] and are therefore effective in optimising the prediction accuracy. Additional examination of Table VI reveals that the performance edge of DL and hybrid models is the greatest at short-term forecasting horizons (0-24 hours), whereby high-frequency temporal patterns can be fruitfully

represented. Conversely, comparative advantage is less stable in medium- and long-term forecasting. To discover more competitive RMSE values like 1.72 in [53] and 0.3909 -0.4582 in [55], simpler ML and simpler DL models can be used, because they need less data and are simpler to model.

The third important point is the excessive focus on very short-term and short-term forecasting horizons. As shown in Fig. 3, 95% (n = 35) of the studies focus on predictions with intervals up to 24 hours ahead, with a range of seconds. Very short-term forecasts of 15 seconds to 1 hour ahead account for 30% (n = 11) of the studies, including [49], [59]-[61], [63], [66]-[69], and [79]. Short-term forecasts of 1 to 24 hours ahead dominate the literature, comprising 65% (n = 24) of studies, including [50]-[52], [54], [56]-[58], [62], [64], [65], [70], [71], and the majority of hybrid studies [72]-[85]. Conversely, the medium- and long-term forecasting horizons are underrepresented, with only 5% (n = 2) of the studies, namely [53] and [55]. This distribution indicates that there is significant research interest in operational forecasting for real-time PV system management, but longer-term forecasting has not been fully explored.

The fourth important observation concerns the prevalence of accuracy-based evaluation measures, especially the RMSE. As shown in Fig. 4, RMSE is the most widely reported metric, appearing in 76% (n = 28) of the studies, including [49]-[54], [56]-[58], [60], [61], [63]-[66], [69]-[76], [78]-[85]. Mean Absolute Error (MAE) is reported in 54% (n = 20) of studies, including [51]-[53], [57], [60], [61], [66], [69], [71], [72], [74]-[76], [78]-[80], [82]-[85]. The coefficient of determination (R^2) appears in 38% (n = 14) of studies, including [53], [54], [58], [60], [63], [66], [70], [72], [74]-[76], [80]-[82]. Mean Absolute Percentage Error (MAPE) and its normalized variants are reported in 32% (n = 12) of studies, including [50], [59], [61], [62], [69], [70], [76]-[79], [81], and [84]. Forecast Skill (FS) is used in 16% (n = 6) of studies, including [49], [56], [73], [75], and [81]. Computational efficiency metrics such as execution time, training time, and resource utilisation are reported in only 11% (n = 4) of the studies, including [49], [50], [65], and [75]. This implies that a great focus is on predictive accuracy, minimal focus on computational performance, and scalability.

The fifth important finding is that there is a low use of multimodal input strategies that combine numerical and visual information. Numerical data, including historical irradiance and meteorological variables, is the most commonly used input type, appearing in 62% (n = 23) of the studies, including [51]-[53], [55], [58], [60], [61], [64], [66], [69]-[74], [76]-[78], [81]-[83] and [85]. A smaller proportion of studies (19%, n = 7) utilise sky or cloud imagery, including [49], [50], [56], [59], [62], [67], and [80]. But only 8% (n=3) of studies present both numeric and pictorial information in a multimodal format, namely, [59], [62], and [80]. This shows that although multimodal learning offers the potential to capture spatial and temporal dynamics, its use remains limited in the existing literature.

The sixth notable point is that comparatively less common is the systematic feature engineering and feature selection methodology. Signal decomposition techniques are used in only 14% (n=5) of the 37 studies and explicit feature selection methods are used in only 16% (n=6). Most research studies do not explicitly specify structured feature engineering pipelines and instead use raw or minimally processed input data. This demonstrates the lack of an opportunity to leverage data preprocessing techniques to improve model performance and interpretability.

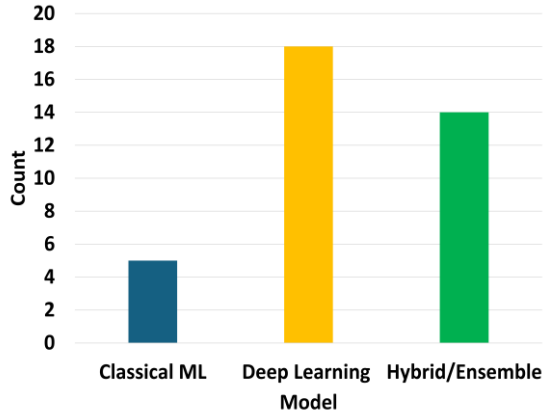


Fig. 2. Distribution of model families across the 37 reviewed studies.

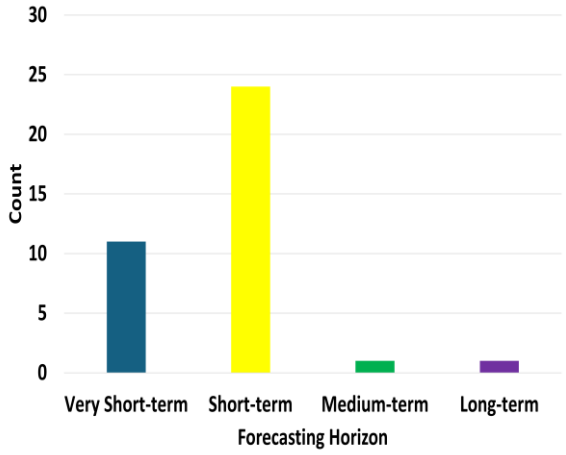


Fig. 3. Distribution of forecasting horizons across the 37 reviewed studies.

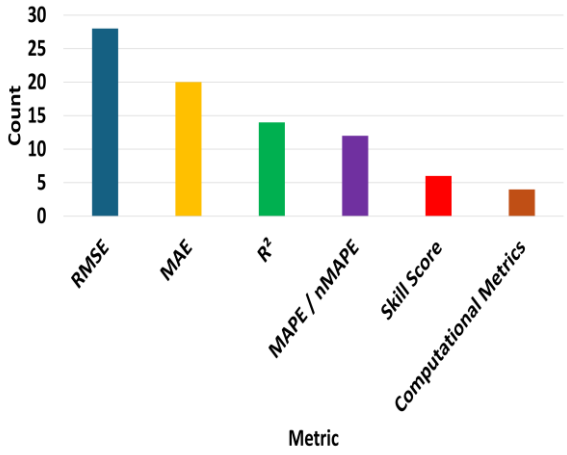


Fig. 4. Frequency of evaluation metrics reported across 37 reviewed studies.

The seventh significant observation is related to the evaluation settings, namely the distribution of single- and multi-

location studies. The majority of the reviewed literature, approximately 65% (n = 24), relies on single-location datasets, including [49]-[53], [56], [58], [59], [61], [63]-[69], [72], [76], [79]-[81], [83], and [84]. In contrast, only 35% (n = 13) of the studies utilise multi-location datasets, such as [54], [55], [57], [60], [62], [70], [73]-[75], [77], [78], [82], and [85]. Even though multi-location studies tend to show better generalisation and model robustness across different climatic conditions, their relatively low use points to a significant gap in the literature. The continued dominance of single-site assessments highlights ongoing issues with data availability and the lack of focus on creating geographically transferable models. Therefore, the applicability of most proposed forecasting methods is questionable, especially across a range of environmental and operating conditions.

All these important observations point to several serious limitations in the existing body of literature, which inform the research gaps considered in the next section.

B. Persistent Limitations and Identified Research Gaps

Although the methodology improved, the overall presentation of the tables shows that there are several crucial and enduring limitations to the successful application and the generalizability of these models.

The first critical gap is the narrowness of the medium-term and long-term forecasting horizons. The literature reviewed is highly biased towards very short and short-term predictions, with 95% (n = 35) of the studies considering forecasting periods of less than 24 hours. Conversely, very few studies 5% (n = 2) discuss the medium- and long-term forecasting. This reveals a major research gap on how to come up with models that can make reliable predictions over a period of days, seasons and over a long period of time, which are crucial in strategic energy planning, grid expansion and policy making.

The second critical gap is that there exists no standardised benchmarking and fair comparison of forecasting models. Even though the DL and hybrid models tend to perform better than traditional ML methods, comparisons are often inconsistent due to differences in datasets, forecasting horizons, input features and evaluation metrics. This complicates the process of coming up with conclusive findings on the most effective models in various circumstances.

The third serious gap is the excessive focus on accuracy-based evaluation measures, with little attention to computational efficiency. Whereas RMSE 76% (n = 28), MAE 54% (n = 20), and R^2 38% (n = 14) appear to be common measures, only 11% (n = 4) of studies report on computational measures such as execution time, training time, and resource utilisation. This means that little attention has been given to model scalability, efficiency, and real-time deployment, especially in resource-limited settings.

The fourth critical gap is the underutilisation of multimodal data integration practices. Even though numerical data dominates in the literature with 62% (n = 23) and only 8% (n = 3) of the studies have both numerical inputs and visual data, like images of the sky. This implies that little attention has been given to the potential of multimodal learning to capture temporal and spatial variations in solar irradiance.

The fifth critical gap is the little use of systematic feature engineering and feature selection methods. Signal decomposition techniques like VMD, CEEMDAN, and wavelet-based signal decomposition techniques are used only in 14% (n = 5) studies, whereas formal feature selection techniques are used in 16% (n = 6) studies. Most studies use raw or minimally processed input data, meaning they do not fully leverage preprocessing strategies that can enhance model performance, robustness and interpretability.

The sixth critical gap is the lack of generalizability as the majority of studies still use datasets from a single location. Even though 59% (n = 22) of the studies use multi-location data, a significant number of 41% (n = 15) of studies remain single-site assessments. This calls into question the strength and applicability of models across other climatic regions and the necessity of more extensive cross-regional validation systems.

The seventh critical gap is the growing sophistication of models with no consideration of interpretability and deployment feasibility. In the literature, there is an increasing tendency to the use of complex DL and hybrid models, such as CNN-LSTM, GAN-based, and transformer models. Nevertheless, these models are seldom tested for their interpretability, computational cost, or practical implementation. This implies that there is a lack of equilibrium between model accuracy, explainability, efficiency and usability in practice.

Overall, the review indicates that the field has made significant progress and has an unmistakably consistent trend towards the more complex hybrid and ensemble deep learning models. This development has, however, also led to an increase in the disparity between methodological innovation and practical applicability. The deep emphasis on gradual advances in predictive accuracy has, in numerous instances, been at the cost of dealing with the underlying issues of real-world implementation. Specifically, poor model generalizability, the lack of adequate focus on computational efficiency and operational constraints, and the continued focus on short-term forecasting horizons are among the barriers. To address these issues, it is necessary to make sure that the models of solar irradiance forecasting have the required operational reliability and practicality in the renewable energy systems.

VI. Recommendations

Based on the critical analysis of the 37 studies and the identified research gaps, several key recommendations are proposed to guide future work toward more robust, generalizable, and deployment-ready solar irradiance forecasting systems.

The high dependence on single-location datasets is one of the significant weaknesses that limit the generalizability of the model to different climatic conditions. Although multi-location studies such as [54], [57], [73], [75], and [82] demonstrate improved robustness, they remain in the minority. Multi-site assessment frameworks based on publicly available data sets like NSRDB, BSRN, ERA5, and NASA POWER should be used in future studies. Specifically, it is suggested to use leave-one-location-out cross-validation to evaluate the spatial robustness. Moreover, domain adaptation methods like feature

alignments and adversarial learning can enhance the transferability of models across climatic regions.

Most analyses focus on predictive accuracy based on measures of RMSE, MAE and R^2, and while computational efficiency is rarely considered, as it is reported in only a small subset of studies, such as [50] and [65]. Future work ought to then take a dual evaluation approach, which takes into account both predictive accuracy and operational performance, such as latency, memory footprint, model size, and energy consumption. This is especially critical with hybrid and DL models, which might not be suitable in real-time or even edge deployment due to the computational bottlenecks.

The growing sophistication of DL and hybrid models requires the implementation of model compression and model optimisation strategies. Pruning, quantisation and knowledge distillation are techniques that can be used to reduce the complexity of models whilst maintaining performance. Moreover, frameworks based on deployment, especially edge-optimised inference pipelines, are to be considered to evaluate the feasibility of microgrid controllers and embedded systems in practice. This is particularly applicable to highly complex models like those reported in studies such as [74], [77] and [83], where deployment feasibility has not been sufficiently addressed.

Literature is extremely biased in regard to very short-term and short-term forecasting horizons, whereas medium-term and long-term forecasting is under-researched. Studies such as [53], [55] and [70] show deterioration in performance with increasing forecasting horizons, and illustrate the challenging nature of long-term prediction in the presence of atmospheric uncertainty. Future studies ought to expand the forecasting horizons to day-ahead and week-ahead by including NWP data, including ERA5 reanalysis data. It is suggested to implement hybrid architectures with CNNs, RNNs, and Transformer-based models to more effectively represent long-range temporal correlations and large-scale weather patterns.

The majority of current research tests models in average conditions without taking into account the variability in atmospheric conditions. Few studies, including [67], evaluate performance under varying conditions of clouds. Future work must then embrace scenario-based evaluation models, in which models are evaluated under clear-sky, partly cloudy, rapidly changing and extreme weather conditions. This stratified assessment is essential to real-world grid integration, where resilience to different conditions is more significant than a mean performance.

There is also a high level of inconsistency in the evaluation methods used in the literature, such as differences in data splitting strategies, normalisation methods, and reported metrics. In spite of the popularity of RMSE in the literature [49]-[85], other similar measures like MAPE, nRMSE, and skill scores are not always reported. Standardisation of benchmarking protocols, such as standard train-validation-test splits, rolling-window or walk-forward validation, and reporting standardised and normalised metrics should be implemented in future studies to increase the reproducibility and comparability of results across different studies.

Lack of standardised benchmark datasets is a significant limiting factor in the field. The available research is based on piecemeal data and does not allow reproducibility and equal comparison. Further work should be done on unified multi-modal benchmark datasets that involve ground-based irradiance measurements, meteorological measurements and satellite or sky-imaging measurements. Partitions of standardised datasets using popular sources (NSRDB, ERA5) should be created, and open benchmarking platforms with leaderboards should be created to speed up the methodological development.

Although the field of DL and hybrid models is improving, there is still a gap between theoretical growth and practical implementation. This is mainly because generalizability, computational efficiency and operational constraints have not been given much consideration. The deployment-conscious model design should thus be the focus of future research where the enhancement in predictive accuracy is matched with scalability, robustness, and computational feasibility.

Overall, the recommendations indicate that the development of models should be no longer based on the predictive accuracy but should be generalizable, operationally robust, and computationally efficient. These gaps need to be solved to bridge the gap between advanced forecasting models in the academic prototyping and their application in real-world energy systems.

VII. Limitations of this SLR

Although this SLR followed PRISMA guidelines, certain limitations remain. The analysis was restricted to English-language publications and selected databases, potentially introducing publication bias. Additionally, the exclusion of grey literature may limit coverage of emerging current, on-the-ground research and industrial practices.

Authors' Contribution

ST Leholo conceptualised the study, developed the methodology, implemented the models, collected and curated the data, performed the formal analysis, and drafted the original manuscript. C Du and T.E Mathonsi contributed to supervision, reviewing and improving the manuscript through proofreading and editorial suggestions. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Tshwane University of Technology.

Acknowledgment

The authors sincerely appreciate the Tshwane University of Technology for providing the necessary support for this research.

References

[1] O. Bamisile, C. Acen, D. Cai, Q. Huang, and I. Staffell, "The environmental factors affecting solar photovoltaic output," Renewable and Sustainable Energy Reviews, vol. 208, no. 115073, pp. 115073-115073, Nov. 2024, doi: https://doi.org/10.1016/j.rser.2024.115073.

[2] S. Poddar, M. Kay, Abhnil Amtesh Prasad, J. P. Evans, and S. Bremner, "Changes in solar resource intermittency and reliability under Australia's

future warmer climate," Solar Energy, vol. 266, pp. 112039-112039, Oct. 2023, doi: https://doi.org/10.1016/j.solener.2023.112039.

[3] O. Gandhi et al., "The value of solar forecasts and the cost of their errors: A review," Renewable and Sustainable Energy Reviews, vol. 189, p. 113915, Jan. 2024, doi: https://doi.org/10.1016/j.rser.2023.113915.

[4] R. J. Jacques Molu et al., "Advancing short-term solar irradiance forecasting accuracy through a hybrid deep learning approach with Bayesian optimization," Results in Engineering, vol. 23, p. 102461, Sep. 2024. doi:10.1016/j.rineng.2024.102461 .

[5] S. Dai et al., "Machine learning in peak demand forecasting: foundations, trends, and insights," Renewable and Sustainable Energy Reviews, vol. 227, p. 116500, Nov. 2025, doi: https://doi.org/10.1016/j.rser.2025.116500.

[6] S. Ghimire, R. C. Deo, D. Casillas-Pérez, and S. Salcedo-Sanz, "Boosting solar radiation predictions with global climate models, observational predictors and hybrid deep-machine learning algorithms," Applied Energy, vol. 316, p. 119063, Jun. 2022, doi: https://doi.org/10.1016/j.apenergy.2022.119063.

[7] M. Attya, O. Abo-Seida, H. Mohamed, and A. Mohammed, "A hybrid deep learning framework for solar irradiation prediction based on regional satellite images and data," Neural Computing and Applications, vol. 37, no. 19, pp. 14327-14363, May 2025, doi: https://doi.org/10.1007/s00521-025-11197-3.

[8] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in Deploying Machine Learning: a Survey of Case Studies," ACM Computing Surveys, vol. 55, no. 6, Apr. 2022, doi: https://doi.org/10.1145/3533378.

[9] S. Liu, D. Lu, S. L. Painter, N. A. Griffiths, and E. M. Pierce, "Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions," Frontiers in Water, vol. 5, Apr. 2023, doi: https://doi.org/10.3389/frwa.2023.1150126.

[10] C. V. Shakila and S. Khadar Babu, "Evaluating and contrasting machine learning and statistical techniques for time series forecasting with hyperparameter optimization," Results in Engineering, vol. 27, p. 106275, Sep. 2025. doi:10.1016/j.rineng.2025.106275.

[11] P. Dubey and M. Kumar, "Integrating Explainable AI with Federated Learning for Next-Generation IoT: A comprehensive review and prospective insights," Computer Science Review, vol. 56, p. 100697, May 2025, doi: https://doi.org/10.1016/j.cosrev.2024.100697.

[12] E. Sarmas, N. Dimitropoulos, V. Marinakis, Z. Mylona, and H. Doukas, "Transfer learning strategies for solar power forecasting under data scarcity," Scientific Reports, vol. 12, no. 1, Aug. 2022, doi: https://doi.org/10.1038/s41598-022-18516-x.

[13] A. Moradzadeh et al., "Generalized global solar radiation forecasting model via cyber-secure Deep Federated Learning," Environmental Science and Pollution Research, vol. 31, no. 12, pp. 18281–18295, Oct. 2023. doi:10.1007/s11356-023-30224-1.

[14] J.U. Omoriare, E.P. Ogherohwo, and J.T. Zhimwang, "Investigating the influence of solar irradiance variability on the output power of photovoltaic (PV) systems in Akure, Nigeria," World Journal of Applied Science & Technology, vol. 16, no. 1, pp. 18-22, May 2025, doi: https://doi.org/10.4314/wojast.v16i1.18.

[15] G. Chesnoiu et al., "Influence of cloudy and clear-sky partitions, aerosols, and geometry on the recent variability in surface solar irradiance components in northern France," Atmospheric Chemistry and Physics, vol. 24, no. 21, pp. 12375-12407, Nov. 2024, doi: https://doi.org/10.5194/acp-24-12375-2024.

[16] R. H. Inman, H. T. C. Pedro, and C. F. M. Coimbra, "Solar forecasting methods for renewable energy integration," Progress in Energy and Combustion Science, vol. 39, no. 6, pp. 535-576, Dec. 2013, doi: https://doi.org/10.1016/j.pecs.2013.06.002.

[17] "Transforming grid operations with accurate short-term energy predictions," DNV, 2024. https://www.dnv.com/article/transforming-grid-operations-with-accurate-short-term-energy-predictions/ (accessed Sep. 19, 2024).

[18] Yuvaraj Mariappan, K. Ramasamy, and Durgadevi Velusamy, "An optimized deep learning based hybrid model for prediction of daily average global solar irradiance using CNN SLSTM architecture," Scientific Reports, vol. 15, no. 1, Mar. 2025, doi: https://doi.org/10.1038/s41598-025-95118-3.

[19] D. T. Mihailović, A. Aksentijevic, and A. Mihailović, "Mapping regularities in the solar irradiance data using complementary complexity measures," Stochastic Environmental Research and Risk Assessment, vol. 35, no. 6, pp. 1257-1272, Dec. 2020, doi: https://doi.org/10.1007/s00477-020-01955-1.

[20] K. Barhmi, C. Heynen, S. Golroodbari, and W. van Sark, "A Review of Solar Forecasting Techniques and the Role of Artificial Intelligence," Solar, vol. 4, no. 1, pp. 99-135, Mar. 2024, doi: https://doi.org/10.3390/solar4010005.

[21] Z. Wang, S. Wu, Y. Huang, R. Liu, and X. Liu, "A comprehensive review of deep learning for solar nowcasting: Enhancing accuracy, reliability, and interpretability," Applied Energy, vol. 407, p. 127378, Mar. 2026, doi: https://doi.org/10.1016/j.apenergy.2026.127378.

[22] X. Li, "Time Series Forecasting with Missing Data Using Generative Adversarial Networks and Bayesian Inference," Information, vol. 15, no. 4, pp. 222-222, Apr. 2024, doi: https://doi.org/10.3390/info15040222.

[23] L. Zhang, R. Wilson, M. Sumner, and Y. Wu, "Transfer learning in very-short-term solar forecasting: Bridging single-site data to diverse geographical applications," Applied Energy, vol. 377, p. 124353, Jan. 2025, doi: https://doi.org/10.1016/j.apenergy.2024.124353.

[24] D. Kibet, M. S. So, and J.-H. Shin, "Minimalist Deep Learning for Solar Power Forecasting: Transformer-Based Prediction Using Key Meteorological Features," Energies, vol. 18, no. 24, p. 6395, Dec. 2025, doi: https://doi.org/10.3390/en18246395.

[25] Md. Z. Hassan, Md. E. Ali, A. B. Ali, and J. Kumar, "Forecasting day-ahead solar radiation using machine learning approach," 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), pp. 252–258, Dec. 2017. doi:10.1109/apwconcse.2017.00050.

[26] O. Trull, J. C. García-Díaz, and A. Peiró-Signes, "A Comparative Study of Statistical and Machine Learning Methods for Solar Irradiance Forecasting Using the Folsom PLC Dataset," Energies, vol. 18, no. 15, p. 4122, Aug. 2025, doi: https://doi.org/10.3390/en18154122.

[27] F. Pandžić and T. Capuder, "Advances in Short-Term Solar Forecasting: A Review and Benchmark of Machine Learning Methods and Relevant Data Sources," Energies, vol. 17, no. 1, p. 97, Dec. 2023, doi: https://doi.org/10.3390/en17010097.

[28] T. Saadati and B. Barutcu, "Forecasting solar energy: Leveraging artificial intelligence and machine learning for sustainable energy solutions," Journal of Economic Surveys, vol. 39, no. 5, pp. 1929–1946, Jan. 2025. doi:10.1111/joes.12678.

[29] R. A. Rajagukguk, R. A. A. Ramadhan, and H.-J. Lee, "A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power," Energies, vol. 13, no. 24, p. 6623, Dec. 2020, doi: https://doi.org/10.3390/en13246623.

[30] K. Razzaq and M. Shah, "Machine Learning and Deep Learning Paradigms: From Techniques to Practical Applications and Research Frontiers," Computers, vol. 14, no. 3, pp. 93-93, Mar. 2025, doi: https://doi.org/10.3390/computers14030093.

[31] A. Takilalte, M. Laissaoui, A. Bouhallassa, L. Mora-López, and S. Harrouni, "Machine learning techniques for forecasting daily global solar radiation ahead," 2023 14th International Renewable Energy Congress (IREC), pp. 1–6, Dec. 2023. doi:10.1109/irec59750.2023.10389357.

[32] A. Javed, B. K. Kasi, and F. A. Khan, "Predicting solar irradiance using machine learning techniques," 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), pp. 1458–1462, Jun. 2019. doi:10.1109/iwcmc.2019.8766480.

[33] M. T. Almuqati, F. Sidi, N. Mohd, M. Zolkepli, and I. Ishak, "Challenges in Supervised and Unsupervised Learning: A Comprehensive Overview," International Journal on Advanced Science Engineering and Information Technology, vol. 14, no. 4, pp. 1449-1455, Aug. 2024, doi: https://doi.org/10.18517/ijaseit.14.4.20191.

[34] S. F. Ahmed et al., "Deep learning modelling techniques: current progress, applications, advantages, and challenges," Artificial Intelligence Review, vol. 56, no. 1, Apr. 2023, doi: https://doi.org/10.1007/s10462-023-10466-8.

[35] X. Hou, C. Ju, and W. Bo, "Prediction of solar irradiance using a convolutional neural network and attention mechanism-based long short-

term memory network based on similar day analysis and an attention mechanism," Heliyon, vol. 9, no. 11, pp. e21484-e21484, Nov. 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e21484.

[36] J. Peter, T. M. Hansen, and H. M. Rekabdarkolaee, "Downscaling and Forecasting Solar Irradiance with RNN, LSTM, and Transformer Models," 2025 57th North American Power Symposium (NAPS), pp. 1-6, Oct. 2025, doi: https://doi.org/10.1109/naps66256.2025.11272413.

[37] A. A. Alguhi and A. M. Al-Shaalan, "LSTM-Based Prediction of Solar Irradiance and Wind Speed for Renewable Energy Systems," Energies, vol. 18, no. 17, pp. 4594-4594, Aug. 2025, doi: https://doi.org/10.3390/en18174594.

[38] M. Ehteram, M. Afshari Nia, F. Panahi, and A. Farrokhi, "Read-First LSTM model: A new variant of long short-term memory neural network for predicting solar radiation data," Energy Conversion and Management, vol. 305, p. 118267, Apr. 2024, doi: https://doi.org/10.1016/j.enconman.2024.118267.

[39] A. Moujahid and F. Dornaika, "Advanced unsupervised learning: a comprehensive overview of multi-view clustering techniques," Artificial Intelligence Review, vol. 58, no. 8, May 2025, doi: https://doi.org/10.1007/s10462-025-11240-8.

[40] J. S. Ramírez and N. D. Duque, "Evaluation of Unsupervised Machine Learning Algorithms with Climate Data," Ingeniería y Desarrollo, vol. 40, no. 02, pp. 131-165, Dec. 2022, doi: https://doi.org/10.14482/inde.40.02.622.553.

[41] E. Obuse et al., "Comparative Analysis of Supervised and Unsupervised Machine Learning for Predictive Analytics," International Journal of Management and Organizational Research, vol. 2, no. 3, pp. 70-86, 2023, doi: https://doi.org/10.54660/ijmor.2023.2.3.70-86.

[42] Khalaf Alsalem, "A hybrid time series forecasting approach integrating fuzzy clustering and machine learning for enhanced power consumption prediction," Scientific Reports, vol. 15, no. 1, Feb. 2025, doi: https://doi.org/10.1038/s41598-025-91123-8.

[43] Y. Zhang, "Enhancing solar irradiance prediction for sustainable energy solutions employing a hybrid machine learning model; improving hydrogen production through Photoelectrochemical device," Applied Energy, vol. 382, p. 125280, Mar. 2025, doi: https://doi.org/10.1016/j.apenergy.2025.125280.

[44] A. Verdone, M. Panella, E. De Santis, and A. Rizzi, "A review of solar and wind energy forecasting: From single-site to multi-site paradigm," Applied Energy, vol. 392, p. 126016, Aug. 2025, doi: https://doi.org/10.1016/j.apenergy.2025.126016.

[45] J.-E. Moussaoui, M. Kmiti, K. El Gholami, and Y. Maleh, "A Systematic Review on Hybrid AI Models Integrating Machine Learning and Federated Learning," Journal of Cybersecurity and Privacy, vol. 5, no. 3, p. 41, Jul. 2025, doi: https://doi.org/10.3390/jcp5030041.

[46] A. Hussain et al., "Hybrid ML/DL Approach to Optimize Mid-Term Electrical Load Forecasting for Smart Buildings," Applied Sciences, vol. 15, no. 18, p. 10066, Sep. 2025, doi: https://doi.org/10.3390/app151810066.

[47] A. R. Singh, R. Seshu Kumar, M. Bajaj, C. B. Khadse, and I. Zaitsev, "Machine learning-based energy management and power forecasting in grid-connected microgrids with multiple distributed energy sources," Scientific Reports, vol. 14, no. 1, Aug. 2024, doi: https://doi.org/10.1038/s41598-024-70336-3.

[48] A. Hakami, "Strategies for overcoming data scarcity, imbalance, and feature selection challenges in machine learning models for predictive maintenance," Scientific Reports, vol. 14, no. 1, p. 9645, Apr. 2024, doi: https://doi.org/10.1038/s41598-024-59958-9.

[49] E. Shirazi, I. Gordon, A. Reinders, and F. Catthoor, "Sky Images for Short-Term Solar Irradiance Forecast: A Comparative Study of Linear Machine Learning Models," IEEE Journal of Photovoltaics, vol. 14, no. 4, pp. 691-698, Jul. 2024, doi: https://doi.org/10.1109/jphotov.2024.3398365.

[50] A. Al-lahham, O. Theeb, K. Elalem, T. A. Alshawi, and S. A. Alshebeili, "Sky Imager-Based Forecast of Solar Irradiance Using Machine Learning," Electronics, vol. 9, no. 10, p. 1700, Oct. 2020, doi: https://doi.org/10.3390/electronics9101700.

[51] B. Alwated and T. Brahimi, "Solar Radiation Forecasting in Saudi Arabia Using Machine Learning," 2021 International Conference on Computers and Automation (CompAuto), pp. 1-6, Sep. 2021, doi: https://doi.org/10.1109/compauto54408.2021.00008.

[52] B. Das, J. L. Cardenas Barrera, and B. Allen, "Solar Energy Forecasting Using Statistical, Machine Learning, and Deep Learning Approaches," 2024 IEEE 12th International Conference on Smart Energy Grid Engineering (SEGE), pp. 177-183, Aug. 2024, doi: https://doi.org/10.1109/sege62220.2024.10739610.

[53] A. Balal, T. Reid, N. Marcheschi, and R. Stewart, "Optimizing Solar Power Forecasting by Employing Various Machine Learning and Deep Learning Strategies for Enhanced Accuracy and Reliability," 2025 IEEE Green Technologies Conference (GreenTech), pp. 46-50, Mar. 2025, doi: https://doi.org/10.1109/greentech62170.2025.10977713.

[54] B. Brahma and R. Wadhvani, "Solar Irradiance Forecasting Based on Deep Learning Methodologies and Multi-Site Data," Symmetry, vol. 12, no. 11, p. 1830, Nov. 2020, doi: https://doi.org/10.3390/sym12111830.

[55] M. Aslam, J.-M. Lee, H.-S. Kim, S.-J. Lee, and S. Hong, "Deep Learning Models for Long-Term Solar Radiation Forecasting Considering Microgrid Installation: A Comparative Study," Energies, vol. 13, no. 1, p. 147, Dec. 2019, doi: https://doi.org/10.3390/en13010147.

[56] C. Feng and J. Zhang, "SolarNet: A sky image-based deep convolutional neural network for intra-hour solar forecasting," Solar Energy, vol. 204, pp. 71-78, Jul. 2020, doi: https://doi.org/10.1016/j.solener.2020.03.083.

[57] O. Bamisile, A. Oluwasanmi, C. Ejiyi, N. Yimen, S. Obiora, and Q. Huang, "Comparison of machine learning and deep learning algorithms for hourly global/diffuse solar radiation predictions," International Journal of Energy Research, vol. 46, no. 8, pp. 10052-10073, Feb. 2021, doi: https://doi.org/10.1002/er.6529.

[58] S. A. Haider, M. Sajid, H. Sajid, E. Uddin, and Y. Ayaz, "Deep learning and statistical methods for short- and long-term solar irradiance forecasting for Islamabad," Renewable Energy, vol. 198, pp. 51–60, Oct. 2022. doi:10.1016/j.renene.2022.07.136.

[59] M. Ajith and M. Martínez-Ramón, "Deep learning based solar radiation micro forecast by fusion of infrared cloud images and radiation data," Applied Energy, vol. 294, p. 117014, Jul. 2021, doi: https://doi.org/10.1016/j.apenergy.2021.117014.

[60] Zaid Allal, H. N. Noura, and K. Chahine, "Machine Learning Algorithms for Solar Irradiance Prediction: A Recent Comparative Study," e-Prime, advances in electrical engineering, electronics and energy, vol. 7, pp. 100453-100453, Mar. 2024, doi: https://doi.org/10.1016/j.prime.2024.100453.

[61] Aneela Zameer, F. Jaffar, F. Shahid, Muhammad Muneeb, R. Khan, and R. Nasir, "Short-term solar energy forecasting: Integrated computational intelligence of LSTMs and GRU," PLOS ONE, vol. 18, no. 10, pp. e0285410-e0285410, Oct. 2023, doi: https://doi.org/10.1371/journal.pone.0285410.

[62] H.-Y. Gao and M. Liu, "Short-term Solar Irradiance Prediction from Sky Images with a Clear Sky Model," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Jan. 2022, doi: https://doi.org/10.1109/wacv51458.2022.00313.

[63] C. N. Obiora, A. N. Hasan, and A. Ali, "Predicting Solar Irradiance at Several Time Horizons Using Machine Learning Algorithms," Sustainability, vol. 15, no. 11, p. 8927, Jan. 2023, doi: https://doi.org/10.3390/su15118927.

[64] S. P. C. Machina, S. S. Koduru, and S. Madichetty, "Solar Energy Forecasting Using Deep Learning Techniques," 2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), Jan. 2022, doi: https://doi.org/10.1109/parc52418.2022.9726605.

[65] Z. Song and L. E. Brown, "Multi-dimensional evaluation of temporal neural networks on solar irradiance forecasting," 2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), pp. 4192–4197, May 2019. doi:10.1109/isgt-asia.2019.8881784.

[66] H. A. Khan, M. Alam, H. A. Rizvi, and A. Munir, "Solar irradiance forecasting using Deep Learning Techniques," IEEC 2023, p. 15, Sep. 2023. doi:10.3390/engproc2023046015.

[67] R. A. Rajagukguk, R. Kamil, and H.-J. Lee, "A Deep Learning Model to Forecast Solar Irradiance Using a Sky Camera," Applied Sciences, vol. 11, no. 11, p. 5049, May 2021, doi: https://doi.org/10.3390/app11115049.

[68] Z. Pang, F. Niu, and Z. O'Neill, "Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons," Renewable Energy, vol. 156, pp. 279–289, Aug. 2020. doi:10.1016/j.renene.2020.04.042 .

[69] A. N.-L. Huynh, R. C. Deo, D.-A. An-Vo, M. Ali, N. Raj, and S. Abdulla, "Near Real-Time Global Solar Radiation Forecasting at Multiple Time-Step Horizons Using the Long Short-Term Memory Network," Energies, vol. 13, no. 14, p. 3517, Jul. 2020, doi: https://doi.org/10.3390/en13143517.

[70] Md. T. Anwar, Md. F. Islam, and Md. G. R. Alam, "Forecasting Meteorological Solar Irradiation Using Machine Learning and N-BEATS Architecture," Proceedings of the 2023 8th International Conference on Machine Learning Technologies, pp. 46-53, Mar. 2023, doi: https://doi.org/10.1145/3589883.3589890.

[71] N. Ghoggali, F. Douak, M. Bensaadallah, and N. Benoudjit, "Comparative Analysis of Machine Learning and Deep Learning Strategies for Solar Irradiation Forecasting," 2023 International Conference on Electrical Engineering and Advanced Technology (ICEEAT), pp. 1-6, Nov. 2023, doi: https://doi.org/10.1109/iceeat60471.2023.10426373.

[72] D. El-Shahat, A. Tolba, M. Abouhawwash, and M. Abdel-Basset, "Machine learning and deep learning models-based grid search cross-validation for short-term solar irradiance forecasting," Journal of Big Data, vol. 11, no. 1, Sep. 2024, doi: https://doi.org/10.1186/s40537-024-00991-w.

[73] P. Kumari and D. Toshniwal, "Extreme gradient boosting and deep neural network-based ensemble learning approach to forecast hourly solar irradiance," Journal of Cleaner Production, p. 123285, Aug. 2020, doi: https://doi.org/10.1016/j.jclepro.2020.123285.

[74] X. Huang et al., "Hybrid deep neural model for hourly solar irradiance forecasting," Renewable Energy, vol. 171, pp. 1041-1060, Jun. 2021, doi: https://doi.org/10.1016/j.renene.2021.02.161.

[75] P. Kumari and D. Toshniwal, "Long short-term memory-convolutional neural network based deep hybrid approach for solar irradiance forecasting," Applied Energy, vol. 295, p. 117061, Aug. 2021, doi: https://doi.org/10.1016/j.apenergy.2021.117061.

[76] N. Elizabeth Michael, M. Mishra, S. Hasan, and A. Al-Durra, "Short-Term Solar Power Predicting Model Based on Multi-Step CNN Stacked LSTM Technique," Energies, vol. 15, no. 6, p. 2150, Mar. 2022, doi: https://doi.org/10.3390/en15062150.

[77] F. Meng et al., "An intelligent hybrid wavelet-adversarial deep model for accurate prediction of solar power generation," Energy Reports, vol. 7, pp. 2155-2164, Nov. 2021, doi: https://doi.org/10.1016/j.egyr.2021.04.019.

[78] Santos et al., "Solar Irradiance Forecasting Using Dynamic Ensemble Selection," Applied sciences, vol. 12, no. 7, pp. 3510-3510, Mar. 2022, doi: https://doi.org/10.3390/app12073510.

[79] Q. Li, D. Zhang, and K. Yan, "A Solar Irradiance Forecasting Framework Based on the CEE-WGAN-LSTM Model," Sensors (Basel), vol. 23, no. 5, pp. 2799-2799, Mar. 2023, doi: https://doi.org/10.3390/s23052799.

[80] Vateanui Sansine, P. Ortega, D. Hissel, and F. Ferrucci, "Hybrid Deep Learning Model for Mean Hourly Irradiance Probabilistic Forecasting," Atmosphere, vol. 14, no. 7, pp. 1192-1192, Jul. 2023, doi: https://doi.org/10.3390/atmos14071192.

[81] P. Singla, M. Duhan, and S. Saroha, "An ensemble method to forecast 24-h ahead solar irradiance using wavelet decomposition and BiLSTM deep learning network," Earth Science Informatics, vol. 15, no. 1, pp. 291-306, Nov. 2021, doi: https://doi.org/10.1007/s12145-021-00723-1.

[82] M. Sivakumar, Jeba Priya S, S. Thomas George, M. S. P. Subathra, Rajasundrapandiyan Leebanon, and Nallapaneni Manoj Kumar, "Nine novel ensemble models for solar radiation forecasting in Indian cities based on VMD and DWT integration with the machine and deep learning algorithms," vol. 108, pp. 108691-108691, May 2023, doi: https://doi.org/10.1016/j.compeleceng.2023.108691.

[83] T. Peng, Y. Li, Z. Song, Y. Fu, Muhammad Shahzad Nazir, and C. Zhang, "Hybrid intelligent deep learning model for solar radiation forecasting using optimal variational mode decomposition and evolutionary deep belief network - Online sequential extreme learning machine," Journal of Building Engineering, vol. 76, pp. 107227-107227, Oct. 2023, doi: https://doi.org/10.1016/j.jobe.2023.107227.

[84] H. Acikgoz, "A novel approach based on integration of convolutional neural networks and deep feature selection for short-term solar radiation forecasting," Applied Energy, vol. 305, p. 117912, Jan. 2022, doi: https://doi.org/10.1016/j.apenergy.2021.117912.

[85] S. Ghimire, B. Bhandari, D. Casillas-Pérez, R. C. Deo, and S. Salcedo-Sanz, "Hybrid deep CNN-SVR algorithm for solar radiation prediction problems in Queensland, Australia," Engineering Applications of Artificial Intelligence, vol. 112, p. 104860, Jun. 2022, doi: https://doi.org/10.1016/j.engappai.2022.104860.

Efficient Computation of Parametric Exponentiation in Parametric Algebra

Khudoykulov Zarifjon[1], Khudoynazarov Umidjon[2], Muminova Mastura[3]
Department of Cryptology, Tashkent University of Information Technologies, Tashkent 100084, Uzbekistan [1]
Department of Cybersecurity & Criminology, Tashkent University of Information Technologies, Tashkent 100084, Uzbekistan [2]
Department of Pedagogy, Fergana State University, Fergana 150100, Uzbekistan [3]

***Abstract*—This study analyzes the computational efficiency of parametric exponentiation operations in parametric algebra and proposes an algorithmic approach for their fast computation. By leveraging the properties of parametric algebra, the possibility of reducing parametric exponentiation to fast exponentiation methods is established, and a corresponding computational algorithm is developed. The proposed approach enables more efficient computation of parametric exponentiation and reduces the computational impact of the system parameter. The obtained results provide a basis for reconsidering approaches to the efficiency of parametric algebra-based computations, demonstrating that the slowness of exponentiation often stems from the computational methods employed rather than the algebraic structure itself. This research establishes a significant foundation for the efficient organization of cryptographic computations based on parametric algebra and for addressing algorithmic optimization challenges.**

***Keywords*—Parametric algebra; parametric multiplication; parametric exponentiation; computational efficiency; algebraic structures; computational complexity; fast exponentiation**

I. Introduction

The security of cryptographic algorithms depends on the computational complexity of the underlying algebraic problems. In symmetric encryption algorithms, complexity is determined by key length, structural design, and resistance to cryptanalytic attacks. In public-key cryptographic algorithms, security primarily relies on the computational complexity of problems such as discrete logarithms in finite fields, integer factorization of large numbers, and computational challenges related to elliptic curves [1, 2].

The rapid development of modern computing technologies is leading to an increase in key lengths and computational complexity to maintain security in existing cryptographic algorithms. However, this approach results in the slowing down of encryption and decryption processes and limits computational efficiency [13].

In recent years, the development of complex cryptographic computing methods based on new algebraic structures has been considered one of the most promising areas. An illustration of this phenomenon is provided by elliptic curve cryptography, which employs more intricate algebraic structures in comparison to classical modular arithmetic. Furthermore, algebraic models such as lattice structures (including LWE), error-correcting codes, and multivariate polynomial equations are actively evolving in modern cryptography. These approaches enrich the mathematical foundations of cryptographic systems and enhance the flexibility of computational processes [21, 22].

The concept of parametric algebra is predicated on the extension of classical and matrix-based algebraic structures through the introduction of an additional parameter. This algebra opens up new mathematical possibilities by generalizing classical algebraic operations through the introduction of an additional parameter. A variety of cryptographic algorithms based on parametric algebra have been proposed and are currently being studied as algebraic models with the potential for application in practical computations [3,4].

Parametric algebra is a system that is only just beginning to gain ground in cryptographic research. Since its structure differs from classical modular arithmetic and it is not yet sufficiently familiar to the wider community, the main concepts and properties are systematically presented in Section III of the study.

Research on parametric algebra has primarily focused on its theoretical properties and general algorithmic applications. However, the issues concerning the computational efficiency of parametric exponentiation operations and the possibilities of adapting them to fast exponentiation methods have not been sufficiently explored.

Existing approaches to parametric exponentiation often rely on repeated parametric multiplications, resulting in high computational overhead and limited practical efficiency. Therefore, developing efficient methods and integrating them with fast exponentiation techniques remains an important scientific problem.

The main contributions of this study can be summarized as follows:

- A novel formulation for parametric exponentiation based on modular binary exponentiation;
- Reduction of the computational dependence on the parameter R in parametric exponentiation;
- Experimental validation using runtime, memory usage, and CPU tick measurements demonstrating efficiency and scalability.

This study focuses on the computational efficiency of parametric exponentiation and its algorithmic optimization.

Leveraging the properties of parametric algebra, a fast modular binary exponentiation algorithm specifically adapted for parametric structures is proposed.

II. Related Works

A. Parametric Algebra and Related Algebraic Approaches

Parametric algebraic operations are based on the concept of generalizing classical algebraic structures through the introduction of an additional parameter. This approach emerges independently across various fields of mathematics and cryptography. This section provides a comprehensive analysis of scientific works that are conceptually and formally related to parametric algebraic operations.

1) Deformations of algebras and parametric operations: In the Theory of Algebraic Deformations developed by Murray Gerstenhaber, an algebraic operation is modified through a parameter [14]. The classical multiplication operation is deformed as follows:

$$a *_t b = ab + t\varphi_1(a,b) + t^2\varphi_2(a,b) \qquad (1)$$

Here, t is the deformation parameter that generates new algebraic structures.

In parametric algebraic operations, classical operations are also modified through a parameter, but over finite fields and by arithmetic expressions. For example:

$$a®b \equiv a + b + a * R * b \ (mod \ \ n) \qquad (2)$$

The similarity between Formula (1) and Formula (2) is that, in both cases, the parameter controls the result of the operation. Therefore, parametric algebra is a discrete, cryptographically oriented variant of deformation theory.

2) Parameter-dependent binary operations: Recent studies show that alternative binary operations are used to construct new mathematical structures. For example, several works on triangular norms have introduced generalized spaces of non-additive measures in which the selected operation determines the resulting functional framework [16].

This illustrates a general principle: modifying or selecting a binary operation can lead to systems with new structural properties. Parametric algebra follows the same direction by introducing parameter-dependent operations over modular and finite arithmetic structures for cryptographic applications. In this sense, it can be regarded as a discrete and cryptography-oriented form of the broader theory of alternative binary operations.

3) Partially associative algebraic structures: In some non-associative algebraic structures, multiplication is governed by specific identities such as:

$$(xy)(xz) = x(yz)x$$

Thus, general associativity is not required, while structural properties are preserved through symmetry [17,18].

In parametric algebraic operations, associativity also depends on the chosen parameter. For certain values of \(R\), the following relation holds:

$$(a®b)®c = a®(b®)$$

This equality is satisfied only for specific parameter values.

Taken together, these studies represent key theoretical frameworks and algorithmic methodologies related to the principles of parametric algebra.

B. Fast Exponentiation and Computational Efficiency

Efficient modular exponentiation methods have been widely studied through several classical approaches. Binary exponentiation is one of the fundamental methods and is applied in left-to-right and right-to-left forms. Its main advantage is reducing the number of multiplications from linear complexity to logarithmic complexity [24].

The Montgomery ladder improves resistance through a regular computation pattern. The sliding-window method increases efficiency by using precomputation. Techniques such as NAF, MOF, and w-NAF reduce the number of nonzero digits in the exponent representation, thereby decreasing the required multiplications [25].

In many public-key systems, the main computational cost is associated with exponentiation operations. In RSA, Diffie–Hellman, digital signature schemes, and related protocols, the exponentiation process has a significant impact on overall efficiency [15].

Recent studies have explored several approaches to address this problem. In particular, classical modular arithmetic methods such as binary, window-based, and other fast exponentiation techniques continue to be improved. In digital signature systems, special attention is also given to accelerating signature generation and verification procedures [19,20].

In specialized algebraic environments, especially Galois fields, methods for optimizing squaring and multiplication operations have been proposed. In addition, new asymptotically faster algorithms are being developed by exploiting modulus factorization or internal algebraic properties of the underlying structures [23].

These results show that improving exponentiation algorithms remains a theoretically and practically relevant research direction today.

Beyond classical algebraic settings, the importance of exponentiation is also significant in parametric algebra. In such systems, the main computational processes are carried out through parameter-dependent addition, multiplication, and exponentiation operations.

Table I presents a comparative summary of the main theoretical approaches related to parametric algebra and the key studies on fast exponentiation.

TABLE I. COMPARATIVE SUMMARY OF RELATED WORKS

No	Ref	General Form of the Algebraic Operation	Algebraic Setting	Main Idea | Relevance	Main Idea | Relevance
1	Gerstenhaber [14]	Related algebraic theory	Deformation theory	Classical algebraic operations are modified through an external parameter, producing new algebraic structures.	Provides a theoretical foundation for parameter-based operations in parametric algebra.
2	Sukhorukova [16]	Parameter-dependent operations	Generalized binary operations	Alternative binary operations generate systems with new structural properties.	Supports redefining arithmetic operations through a parameter.
3	Grishkov and Zavarnitsine [17]	Non-associative structures	Groups with triality	Structural identities may replace classical associativity while preserving algebraic consistency.	Related to parameter-dependent structural behavior.
4	Hall [18]	Non-associative structures	Moufang loops and triality	Studies algebraic systems where symmetry and special identities replace standard associativity.	Conceptually related to generalized algebraic operations.
5	Marouf et al. [24]	Fast exponentiation methods	Modular arithmetic	Comparative study of binary methods, Montgomery ladder, sliding-window, NAF, MOF, and w-NAF.	Provides classical foundations of efficient exponentiation methods.
6	El Makkaoui et al. [20]	Fast exponentiation methods	Public-key cryptography	Reviews practical modular exponentiation methods for cryptographic systems.	Shows the practical importance of efficient exponentiation.
7	El Makkaoui et al. [19]	Cryptographic optimization	RSA cryptosystem	Fast RSA variants reduce the cost of decryption and digital signature generation.	Confirms exponentiation efficiency remains critical in real systems.
8	Doukas [23]	Specialized exponentiation	Galois fields GF(2^n)	Fast squaring, efficient multiplication, and precomputation significantly accelerate exponentiation.	Demonstrates optimization in alternative algebraic environments.
9	Aggarwal and Isaacs [15]	Asymptotic improvement	Factored modulus arithmetic	Faster modular exponentiation is achieved by exploiting factorized modulus structure.	Indicates exponentiation remains an active theoretical research topic.
10	Somsuk et al. [25]	Parallel acceleration	Modular arithmetic	Improved modular exponentiation using periodic bit-pattern exponents and acceleration techniques.	Shows that implementation-level optimization is still actively studied.

The efficiency of parametric exponentiation directly affects the practical performance of these schemes. Therefore, when parametric algebra is applied in cryptographic systems, not only security but also the computational efficiency of algebraic operations becomes important. This requires a separate consideration of existing parametric cryptosystems and the role of parametric exponentiation within them. The next section is devoted precisely to these issues.

C. Existing Parametric Cryptosystems and Research Gap

Several cryptographic algorithms have been proposed based on parametric algebra, and some of them have reached the level of standardization. Symmetric, public-key, hash-function, and digital signature systems enhanced by parametric algebra are presented in Fig. 1. The figure also shows the hierarchical use of parametric operations in key generation, encryption or signing, and decryption or verification stages [5-11].

As shown in Fig. 1, parametric multiplication, parametric inverse operations, and parametric exponentiation are used as core computational components in different cryptographic mechanisms. In particular, parametric exponentiation appears in several stages of public-key systems and digital signature schemes, which highlights its practical importance.

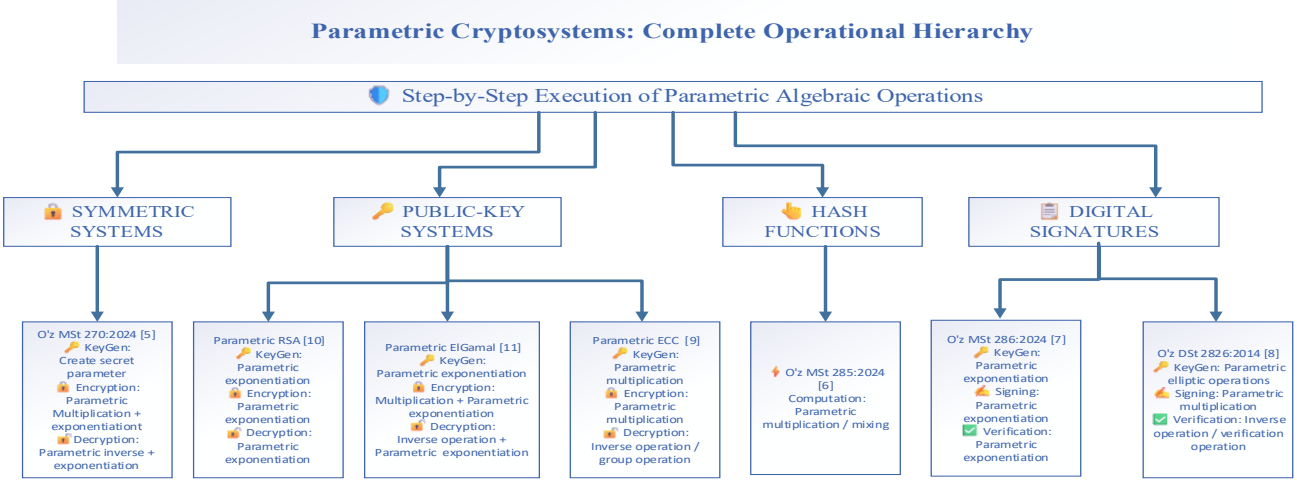


Fig. 1. Parametric cryptosystems

Existing parametric cryptosystems show that parametric exponentiation is a central operation in many schemes. However, in most cases it is computed by iterative or sequential procedures based on repeated parametric multiplications. This leads to increased computational cost, especially for large exponents and parameter values, motivating the need for more efficient methods.

As a result, expanding the practical applicability of parametric algebraic systems requires optimization of their core operations, especially exponentiation. Therefore, developing efficient parametric exponentiation methods remains an important research problem.

III. Problem Statement

In parametric algebraic systems, one of the main operations is parametric multiplication, which is defined as follows:

$$a®b \equiv a + b + a * R * b \ (mod \ \ n),$$

where, R is an additional parameter of the algebraic structure. Based on this operation, parametric exponentiation is widely used in existing cryptographic algorithms. In practice, it is usually represented in the following form:

- Initial condition: $a^{\backslash 1} = a \ mod \ n$.
- Recursive step: $a^{\backslash x+1} = a + a^{\backslash x}(1 + Ra) \ mod \ n$

A direct algorithm for computing this expression can be written as follows (see Algorithm 1):

Algorithm 1: Existing Iterative Method for Parametric Exponentiation

```
Input: Integers a, k, R, n
Output: a^(k) mod n
1: if k = 1 then
2: return a mod n
3.  end if
4. t ← (1 + R·a) mod n
5: for i = 2 to k do
6:     result ← (a + result·t) mod n
7:     end for
8: return result
```

This algorithm is a simple and direct computation method. However, the parametric multiplication operation is performed sequentially $k - 1$ times. Therefore, the computational cost increases as the exponent value grows.

In addition, since all operations depend on the parameter R, the execution time may become more significant for large parameter values.

Several methods exist for computing parametric exponentiation in systems based on parametric algebra. These include recursive iterative methods, sequential procedures based on repeated parametric multiplication, and summation or closed-form formulas. However, practical computational efficiency remains an important issue.

Thus, the main problem is to develop a more efficient method for computing parametric exponentiation while preserving the correctness of the algebraic system.

IV. Parametric Algebra and its Properties

Since parametric algebra is a relatively new algebraic framework in cryptographic research and remains less familiar to the broader community, its main operations, fundamental properties, and illustrative examples are presented in this section.

Parametric algebra extends classical algebraic operations by introducing an additional parameter. This parameter increases computational complexity and thereby enhances cryptographic resilience. Consequently, it provides sufficient security for encryption algorithms based on parametric algebra in large finite fields.

Parametric algebra proposes an extended form of classical arithmetic operations, serving as a promising foundation for improving the mathematical structure of cryptographic algorithms and incorporating an additional complexity component [3].

In *parametric algebra*, the parameter R is selected from the set of numbers in the field $Z_p = \{0,1, \dots, n-1\}$ and all operations are performed modulo n.

A. Multiplication with Parameter R

Definition 1. For any $a, b \in Z$, the expression:

$$a®b \equiv a + b + a * R * b \ (mod \ \ n)$$

defines the *parametric multiplication operation* with parameter R. In the context of parametric algebra, the symbol R is referred to as either a *coefficient* or a *transformation parameter*. When $R = 0$, this expression reduces to the *classical algebraic addition* operation [3].

Example: If $a = 3, b = 5, R = 2, n = 17$ then,

$$3 \ ® \ 5 \ mod \ n = 3 + 5 + 3 \cdot 2 \cdot 5 \ mod \ 17 = 38 \ mod \ 17 = 4$$

This operation serves as the fundamental functional unit of the proposed parametric algebraic structure.

The parametric multiplication operation defined above satisfies the following properties:

Closure. For arbitrary $a, b \in Z_n$, the operation is closed:

$$a \ ® \ b = c \ mod \ n, c \in Z_n$$

Commutativity. The operation is commutative, as $a \ ® \ b = b \ ® \ a \ mod \ n$ for all $a, b, R \in Z_n$. This result follows from the symmetry of the given expression:

$$a + b + a \ \cdot R \ \cdot b = a + b + \ b \cdot R \ \cdot a$$

Associativity. The operation is associative for all $a, b, R \in Z_n$: $(a \ ® \ b)® \ c \ mod \ n = a \ ® \ (b \ ® \ c) mod n$.

Identity Element. There exists an identity element $e = 0$, where $a®e = a + 0 + a \cdot R \cdot 0 = a \ mod \ n$ for any $a, e, R \in Z_n$

Inverse Element. For each $a \neq 0$ such that $\gcd(a + aR, n) = 1$, there exists an inverse element a^{-1} where $a \ ® \ a^{-1} = e \ mod \ n$.

B. Inversion Operation in Parametric Algebra

Parametric inversion operation modulo n with parameter R is defined as:

$$a^{\backslash -1} \equiv -a * (1 + R * a)^{-1} \ (mod\ n)$$

Here, the notation $^{-1}$ denotes the classical modular inversion. The element $^{\backslash -1}$ represents the parametric inverse with respect to the parameter R and modulus n and satisfies the identity $a \circledR a^{-1} \equiv 0 (mod\ n)$ [26].

Verification.

To verify this, we substitute the expression $a^{\backslash -1} \equiv -a * (1 + R * a)^{-1} \ (mod\ n)$ into the parametric multiplication formula $a \circledR a^{\backslash -1} = a + a^{\backslash -1} + a \cdot R \cdot a^{\backslash -1} mod\ n$:

$a \circledR a^{\backslash -1} mod\ n = a + (-a(1 + R * a)^{-1}) + a \cdot R \cdot (-a(1 + R * a)^{-1})\ mod\ n = a - a(1 + R * a)^{-1} - a \cdot R \cdot a(1 + R * a)^{-1}\ mod\ n = a - a(1 + R * a)^{-1}(1 + R * a)\ mod\ n$

Since $-a(1 + R * a)^{-1}(1 + R * a)\ mod\ n = 1$, we obtain $a - a \equiv 0\ mod\ n$.

Example: Let $a = 4, R = 2, n = 17$. First, we calculate the parametric inverse using the formula $a^{\backslash -1} \equiv -a * (1 + R * a)^{-1} \ (mod\ n)$:

Calculate $1 + aR = 1 + 4 \cdot 2\ mod\ 17 = 9$ mod 17

Find the modular inverse: $(1 + R * a)^{-1}\ mod n = 9^{-1} mod\ 17 = 2\ mod\ 17$

Compute the final result: $a^{\backslash -1} = -a * (1 + R * a)^{-1} = -4 \cdot 2 = -8\ mod\ 17 = 9\ mod\ 17$

Verification: $a \circledR a^{-1} mod\ n = 4 \circledR 9\ mod\ 17 = 4 + 9 + 4 \cdot 2 \cdot 9\ mod\ 17 = 13 + 72\ mod\ 17 = 85\ mod 17 = 0$

The inversion operation defined above establishes a robust algebraic framework. The following properties define the order and cardinality of the resulting parametric multiplicative group:

Property 1. If the modulus n is a prime number and the parameter R of the parametric multiplicative group is any natural number, then the order of the group is equal to $\varphi(n)$, where $\varphi(n)$ represents the value of *Euler's totient function*.

Property 2. For a parametric multiplicative group with a composite modulus n if the parameter R is coprime to the modulus $(\gcd(R, n) = 1)$, then the order of the group is equal to $\varphi(n)$, where $\varphi(n)$is the value of Euler's totient function.

Implications for Inversion Operations. These properties are critical for the inversion process. The group order $\varphi(n)$$ ensures that for every valid element, a unique inverse exists within the structure. Furthermore, according to the parametric adaptation of Euler's Theorem, the inverse can also be computed through exponentiation:

$$a^{-1} = a^{\backslash \varphi(n) - 1}\ mod\ n$$

This dual path for inversion enhances the cryptographic flexibility of the proposed system.

C. Parametric Exponentiation Operation

The concept of parametric exponentiation is based on the repeated application of the parametric multiplication operation.

Definition 2. In a parametric multiplicative group modulo n with a parameter $R \geq 1$, the operation of exponentiation is referred to as a parametric exponentiation function.

The result of raising a base a to the power x is defined by the following recursive relation:

Initial condition: $a^{\backslash 1} = a\ mod\ n$.

Recursive step: $a^{\backslash x+1} = a + a^{\backslash x}(1 + Ra)\ mod\ n$

where, $x \in N$ and the element $^{\backslash 1}$ denotes the parametric exponentiation operator [3].

Example. Let $n = 17, R = 2, a = 3$ and $k = 4$. We aim to calculate the parametric exponentiation value $y = 3^{\backslash 4}\ mod 17$.

First, calculate the transform factor F:

$F = 1 + R \cdot a\ mod\ n = 1 + 2 \cdot 3\ mod 17 = 7\ mod 17$

Next, the sequence is recursively defined by $d_1 = a, d_{t+1} = a + F \cdot d_t\ mod\ n$.

The sequence d_t represents the iterative computation of parametric exponentiation.

$d_1 = a = 3,$

$d_2 = a + Fd_1\ mod\ n = 3 + 7 \cdot 3 = 3 + 21\ mod 17 = 7\ mod 17$

$d_3 = 3 + 7 \cdot 7 = 3 + 49\ mod 17 = 1\ mod 17$

$d_4 = 3 + 7 \cdot 1 = 3 + 7\ mod 17 = 10\ mod 17$

The iterative calculation yields $3^{\backslash 4}\ mod 17 = 10\ mod 17$, which confirms the result of the parametric power function for the given parameters.

The parametric exponentiation function exhibits several fundamental properties as outlined below [3].

Property 1. For arbitrary $a, z, d \in \mathbb{Z}_n$, the following relations hold:

$a^{\backslash z+d} = a^{\backslash z} \circledR a^{\backslash d} mod\ n,\ a^{\backslash z} = a^{\backslash z} \circledR e\ mod\ n$

where, $\circledR -$ denotes the parametric multiplication operation modulo n, e is the identity element and $a^{\backslash x}$ - denotes the parametric exponentiation operator.

Property 2. For arbitrary $a, z, d \in \mathbb{Z}_n$, the following relations hold: $a^{\backslash z*d} = (a^{\backslash z})^{\backslash d} = (a^{\backslash d})^{\backslash z}\ mod\ n$

where, $a^{\backslash x}$ - denotes the parametric exponentiation operator.

Property 3. For any $a \in \mathbb{Z}_n$, the following holds:

$a^{\varphi(n)+1} = a\ mod\ n,\ a^{\backslash 1} = a$, where $\varphi(n)$denotes Euler's totient function.

Property 4. If d and e are coprime to $\varphi(n)$ and satisfy the condition $d \cdot e \equiv 1\ mod\ (\varphi(n))$, then the following relation holds: $(a^{\backslash d})^{\backslash e} = a\ mod\ n$, where $a \in \mathbb{Z}_n$ and $a^{\backslash x}$ - denotes parametric exponentiation operator.

Property 5. For any $a \in \mathbb{Z}_n$, the following relations hold:

$$a^{\backslash 0} = 0, a^{\backslash -1} = -a * (1 + R * a)^{-1} \ mod\, n$$

where, $a^{\backslash -1} -$ represents the parametric inverse of a modulo n and $(1 + R * a)^{-1}$ denotes the classical modular inverse of $(1 + R * a)$ modulo n.

Property 6. The relationship between parametric exponentiation and classical modular exponentiation is given by

$$R * a^{\backslash x}\ (mod n) = (1 + R * a)^x - 1\ mod\ n$$

where R is a parameter and $\neq 0$, $\in \{1,2,.\varphi(n) - 1\}$, $a \in \{1,2,..,n-1,$ $a^{\backslash x}$ - denotes parametric exponentiation operator.

Property 7. Let $F = 1 + Ra$. Then, for parametric exponentiation,

$$a^{\backslash x+1} \equiv a * \sum_{i=0}^{i=x} F^{\,i} (mod\ n)$$

where, R is a parameter, $a^{\backslash x+1}$ - denotes the parametric exponentiation operator, $\sum_{i=0}^{i=x} F^{\,i} (mod\ n)$ denotes the geometric sum.

The algebraic properties presented above facilitate the enhancement of existing cryptographic algorithms and enable more efficient computational processes within parametric structures.

For a more comprehensive discussion and additional definitions of parametric algebraic operations, the reader is referred to [3].

V. Efficiency Analysis of Parametric Exponentiation

In this section, the computational efficiency of parametric exponentiation is analyzed. Various approaches, including iterative and closed-form methods, are examined. In addition, methods for improving algorithmic efficiency are proposed by applying classical modular binary exponentiation techniques to existing algorithms.

Several publications and papers on parametric algebra present both iterative (sequential) and closed-form methods for exponentiation with parameter R [3, 4, 12].

A. Iterative Parametric Exponentiation

Let the parametric multiplication operation modulo n be defined as follows:

$$a®b \equiv a + b + a * R * b\ (mod\ \ n),$$

Where $R \in Z_n$ is a parameter, $a, b \in Z_n$. This operation differs from classical multiplication and represents a combination of addition and multiplication.

The exponentiation operation based on this parametric multiplication is defined as:

$$a^{\backslash k} = \underbrace{a®a®a® \ldots . a®}_{k\ times}$$

Specifically, $a^{\backslash 1} = a$, $a^{\backslash k+1} = a^k ® a$. This expression is an analogous form of the classical exponentiation operation, where the parametric multiplication is utilized instead of the standard multiplication.

Parametric exponentiation satisfies the following recursive relation:

$$a^{\backslash k+1} = a^{\backslash k} ® a$$

Using the definition of the parametric multiplication, we obtain $a^{\backslash k+1} = a^{\backslash k} + a + a^{\backslash k} \cdot R \cdot a\ mod n$

This expression can be rewritten as follows:

$$a^{\backslash k+1} = a + a^{\backslash k}(1 + Ra) mod\ n$$

Let $d_k = a^{\backslash k}$ Then, the recurrence relation can be written as

$$d_{k+1} = a + d_k(1 + Ra) mod\ n \quad (3)$$

This recurrence defines the iterative computation of parametric exponentiation up to the exponent k.

By leveraging the algebraic properties of parametric exponentiation, the recurrence relation, Formula (3), can be transformed into equivalent alternative representations. These forms facilitate further optimization and significantly enhance computational efficiency.

B. Alternative Forms of Parameteric Exponentiation

Based on Property 7 of parametric exponentiation, the following *summation-based representation* is derived:

$$a^{\backslash x+1} \equiv a * \sum_{i=0}^{i=x} F^{\,i} (mod\ n). \quad (4)$$

Based on Formula (4), the exponentiation operation can be computed without relying on repeated squaring, by evaluating a geometric series and performing a single modular multiplication with the base a. This approach reduces the number of modular multiplications, enables efficient precomputation, and eliminates bit-dependent branching. Consequently, it yields a more efficient exponentiation mechanism for cryptographic algorithms based on parametric algebra.

The closed-form expression Formula (4) can be *derived from the recurrence relation* Formula (3) as follows:

According to Formula (3), we have $d_1 = a$, $d_2 = a + d_1(1 + Ra)$, $d_3 = a + d_2(1 + Ra)$,…

To simplify the expression, the following notation is introduced:

$$F = 1 + Ra$$

Then, the recurrence relation can be written as

$$d_{k+1} = a + F \cdot d_k mod\ n$$

Expanding this recurrence iteratively yields:

$$d_1 = a$$

$$d_2 = a + Fa$$

$$d_3 = a + F(a + Fa) = a + Fa + Fa^2$$

$$d_4 = a + F(a + Fa + Fa^2) = a + Fa + F^2a + F^3a$$

Continuing this pattern, the general term for d_k is expressed as:

$$d_k = a(1 + F + F^2 + \cdots + F^x)$$

Consequently,

$$a^{\backslash k+1} = a\sum_{i=0}^{k} F^i \ mod\ n, F = 1 + Ra$$

This representation corresponds exactly to Formula (4), derived from *the 7th property* of parametric exponentiation.

Based on the 6th property of parametric exponentiation:

$$R * a^{\backslash k}\ (mod n) = (1 + R * a)^k - 1\ mod\ n$$

it is possible to derive the *closed-form expression* for parametric exponentiation:

$$a^{\backslash k} = ((1 + Ra)^k - 1) \cdot R^{-1}\ mod\ n \qquad (5)$$

In this context, it is necessary to satisfy the condition $\gcd(R, n) = 1$ to ensure the existence of the modular inverse of the parameter R.

The derivation of the closed-form Formula (5) from the recurrence relation Formula (3) is expressed as follows.

$$\text{Let } F = 1 + Ra.$$

Substituting this into Formula (3), the recurrence relation takes the following form:

$$d_{k+1} = a + F \cdot d_k mod\ n$$

Expanding this recurrence iteratively yields:

$$d_1 = a$$

$$d_2 = a + Fa$$

$$d_3 = a + F(a + Fa) = a + Fa + Fa^2$$

$$d_4 = a + F(a + Fa + Fa^2) = a + Fa + F^2a + F^3a$$

$$.d_k = a(1 + F + F^2 + \cdots + F^{k-1}) mod\ n$$

This sequence represents a geometric progression. From standard mathematics, the sum of a geometric series is given by:

$$x + x \cdot q + x \cdot q^2 + \cdots + x \cdot q^{k-1} = x \cdot \frac{q^k - 1}{q - 1}$$

In our case, $x = a, q = 1 + Ra$. Substituting these values, we derive:

$$d_k = a \cdot \frac{(1 + Ra)^k - 1}{(1 + Ra) - 1} = a\frac{(1 + Ra)^k - 1}{Ra}$$

$$= ((1 + Ra)^k - 1)\ R^{-1}\ mod\ n$$

Hence,

$$a^{\backslash k} = ((1 + Ra)^k - 1) \cdot R^{-1}\ mod\ n$$

This coincides with the closed-form expression given in Formula (5).

Although the resulting parametric exponentiation expressions provide non-iterative closed-form representations, they still require further refinement to achieve high computational efficiency. Therefore, additional optimization techniques are necessary to further accelerate their evaluation.

Efficient exponentiation techniques play a crucial role in modern cryptographic computations. Among them, binary exponentiation is one of the most widely used methods due to its logarithmic complexity.

C. Efficient Computation Using Binary Exponentiation

For given $a, k, n \in \mathbb{N}$, one of the most efficient methods for computing $a^k mod n$ is binary exponentiation. This method is based on representing the exponent k in its binary form:

$$k = \sum_{i=0}^{t} b_i 2^i,\ b_i \in \{0,1\}$$

Accordingly, the exponentiation can be expressed as

$$a^k = \prod_{i=0}^{t} a^{b_i 2^i}$$

The computation is performed by squaring and multiplying successively, with modular reduction applied at each step to prevent the intermediate values from becoming too large.

The pseudocode for this algorithm is as follows (see Algorithm 2):

Algorithm 2: Binary Modular Exponentiation

```
Input: Integers a, k, n
Output: a^k mod n
1: result ← 1
2: base ← a mod n
3: while k > 0 do
4:    if k mod 2 = 1 then
5:          result ← (result · base) mod n
6:    end if
7:    base ← (base · base) mod n
8:    k ← ⌊k / 2⌋
9: end while
10: return result
```

The algorithm relies on the binary representation of the exponent k, enabling efficient computation through successive squaring and conditional multiplication. Modular reduction at each step prevents the growth of intermediate values. With a computational complexity of $O(\log k)$, this method is widely regarded as one of the most efficient techniques in cryptographic applications.

D. Integrating Binary Algorithms into Parametric Algebra

The binary exponentiation technique can be applied to different representations of parametric exponentiation. In particular, it can be integrated with both the iterative formulation Formula (3) and the closed-form representation Formula (5).

The *recursive form of parametric* exponentiation is given by Formula (3):

$$d_{k+1} = a + d_k(1 + Ra)(\text{mod}n).$$

This formulation enables a sequential computation of parametric exponentiation. However, it exhibits linear time complexity in the exponent k, namely $O(k)$.

To improve computational efficiency, the classical binary exponentiation technique can be applied to the given recurrence

process. In this approach, the exponent k is represented in its binary form:

$$k = \sum_{i=0}^{t} b_i 2^i, b_i \in \{0,1\}.$$

Based on this representation, parametric exponentiation can be expressed as:

$$a^{\backslash k} = \underbrace{a®a®a® \ldots . a®}_{k\ times}$$

where, the parametric multiplication is defined as:

$$a®b = a + b + aRb(\mathrm{mod}n).$$

The computation follows the structure of binary exponentiation, involving successive squaring and conditional multiplication according to the binary representation of the exponent**.**

As a result, the following iterative algorithm is obtained. If the current bit of k is equal to 1, the result is updated using parametric multiplication, while at each step the base is updated through parametric squaring.

This approach reduces the number of computational steps to $O(\log k)$.

Therefore, the application of binary exponentiation significantly reduces the number of iterations required for computing parametric exponentiation.

The pseudocode for this algorithm is as follows (see Algorithm 3):

Algorithm 3: Iterative Binary Parametric Exponentiation

```
Input: Integers a, k,R, n
Output: a^k mod n
1: result ← 0
2: base ← a mod n
3: while k > 0 do
4:    if k mod 2 = 1 then
5:          result ← (result + base + R · result · base) mod n
6:    end if
7:    base ← (base + base + R · base · base) mod n
8:    k ← ⌊k / 2⌋
9: end while
10: return result
```

Example. Let, $R = 19$, $a = 5$, and $k = 13$.

We compute the value of parametric exponentiation.

Binary representation:

$$k = 13 = (1101)_2$$

The parametric multiplication is defined as

$$a®b = a + b + aRb(\mathrm{mod}n).$$

Initial values:

$$\mathrm{result} = 0, \mathrm{base} = 5.$$

Step 1 ($k = 13$, bit = 1)

$$\mathrm{result} = 0®5 = 5$$

$$\mathrm{base} = 5®5 = 5 + 5 + 19 \cdot 25 = 485 \equiv 23(\mathrm{mod}77)$$

Step 2 ($k = 6$, bit = 0)

$$\mathrm{result} = 5$$

$$\mathrm{base} = 23®23 = 23 + 23 + 19 \cdot 529 = 10097 \equiv 10(\mathrm{mod}77)$$

Step 3 ($k = 3$, bit = 1)

$$\mathrm{result} = 5®10 = 5 + 10 + 19 \cdot 50 = 965 \equiv 41(\mathrm{mod}77)$$

$$\mathrm{base} = 10®10 = 10 + 10 + 19 \cdot 100 = 1920 \equiv 72(\mathrm{mod}77)$$

Final result:

$$a^{\backslash 13} \equiv 68(\mathrm{mod}77).$$

Formula (5), derived from Property 6 of parametric exponentiation, transforms the parametric exponentiation into a classical modular exponentiation form. In this formulation, the main computational cost is associated with evaluating:

$$F^k = (1 + Ra)^k.$$

Since this expression corresponds to classical exponentiation, it can be efficiently computed using standard modular exponentiation techniques.

Thus, parametric exponentiation can be reduced to a classical modular exponentiation problem via Formula (5), where the value F^k is efficiently computed using binary exponentiation.

The pseudocode for computing parametric exponentiation using the binary exponentiation method is as follows (see Algorithm 4):

Algorithm 4: Binary Parametric Exponentiation Closed-form

```
Input: Integers a, k, R, n
Output: a^k mod n
1: if k = 0 then
2: return 0
3: while k > 1 do
4:    A ← (1 + R · a)mod n
5:    invR ← R^{-1}modn  //assuming gcd(R,n)=1
6:    t ← BinaryExponantiation (A, k, n)
7:    result ← (t − 1)mod n
8:    result ← (result · invR)modn
9: return result
```

Example. Given the values $n = 77, R = 19$, $a = 5$, and $k = 13$, compute $a^{(k)}\mathrm{mod}\ n$ using the closed-form representation Formula (5).

Step 1. Using Formula (5)

$$a^{\backslash k} = ((1 + Ra)^k - 1) \cdot R^{-1}\ mod\ n$$

Step 2. Compute $A = 1 + Ra\mathrm{mod}\ n$

$$A = 1 + 19 \cdot 5\ mod\ 77 = 19\ mod\ 77$$

Step 3. Compute the modular inverse of R

$$R^{-1} = 19^{-1} \equiv 73 (\mathrm{mod} 77).$$

Step 4. Compute 19^{13} mod 77using binary exponentiation.

$$13 = 1101_2 = 2^3 + 2^2 + 2^0$$

$$19^2 \equiv 53, 19^4 \equiv 37, 19^8 \equiv 60 (\mathrm{mod} 77)$$

$$19^{13} \equiv 19^8 \cdot 19^4 \cdot 19 \equiv 60 \cdot 37 \cdot 19 \equiv 61 (\mathrm{mod} 77)$$

Step 5. Final computation

$$a^{(13)} = (61 - 1) \cdot 73 = 60 \cdot 73 = 4380 \equiv 68 (\mathrm{mod} 77).$$

The analysis of the presented example demonstrates that the considered parametric exponentiation algorithms, based on Formula (3) and Formula (5), improve computational efficiency compared to the direct iterative approach.

Simple iterative parametric exponentiation method requires 13 parametric multiplications; both algorithms leverage binary exponentiation to reduce the number of iterations to ($\log k$).

Moreover, the approach based on the closed-form representation Formula (5) further simplifies the computation by reducing it to a single modular exponentiation, resulting in improved practical efficiency.

The classical modular exponentiation and the parametric methods based on Formula (3) and Formula (5) differ significantly in their algebraic foundations and computational processes. A generalized comparison of the characteristics of these exponentiation algorithms is presented in Table II.

TABLE II. PARAMETRIC CRYPTOGRAPHIC ALGORITHMS

Property	Classical	Formula (3)	Formula (5)
Algebraic basis	Classical modular arithmetic	Parametric algebra	Parametric algebra
Main operation	Modular multiplication	Parametric multiplication	Modular multiplication
Dependence on parameter	None	Strong	Weak
Computational process	Squaring and multiplication	Iterative parametric multiplication	Binary exponentiation
Number of parametric operations	0	≈ k times	0
Algorithmic complexity	O(log k)	O(k)	O(log k)

VI. EXPERIMENTAL RESULTS AND DISCUSSION

To improve the efficiency of parametric exponentiation, the fast binary exponentiation technique was applied to the parametric exponentiation methods considered in this study. The performance of the considered algorithms was evaluated using three metrics: *execution time*, *memory consumption*, and *CPU ticks*, in order to provide a broader view of their practical efficiency.

The performance of the considered exponentiation algorithms was evaluated with respect to the modulus $n$and the parameter R, in order to analyze their impact on computational efficiency.

To this end, software implementations of three algorithms were developed: the *classical* modular exponentiation algorithm, the *iterative* parametric exponentiation method based on Formula (3), and the optimized *closed-form* parametric exponentiation algorithm based on Formula (5).

For clarity, these algorithms are denoted as *Classic*, *Param5*, and *Param7*, respectively.

The software modules were implemented in the C# programming language and executed on the .NET platform.

To ensure the reliability of the performance metrics, the experiments were conducted on a computing system with the following technical specifications: Windows 11 Pro operating system, Intel Core i7-13700U processor with a clock frequency of 2.40 GHz, and 16 GB of RAM.

To ensure the reliability of the results, each algorithm was executed multiple times for each set of parameter values. The exponentiation operation was repeated 100 times within the program and the execution time was measured using a high-precision timer.

A. Impact of the Modulus n

The performance of the algorithms was evaluated for different lengths of the modulus n. The parameters used in the computations were fixed as R=2048 bit, a=512 bit and k=256 bit.

The modulus length was gradually increased starting from 512 bits, and for each tested value, the algorithms were evaluated in terms of execution time, memory consumption, and CPU ticks.

The obtained results, illustrating the dependence of execution time on the modulus bit-length, are presented graphically in Fig. 2.

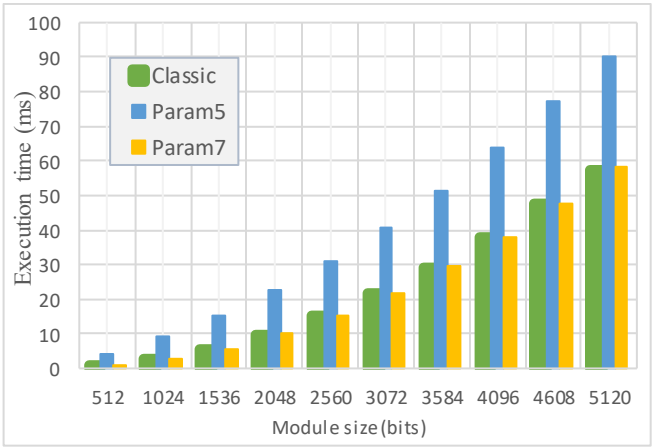


Fig. 2. Execution time vs. modulus length.

The graphical results show that the execution time of all algorithms increases as the modulus bit-length grows. The iterative parametric exponentiation method based on Formula (3) operates more slowly than the classical algorithm, since the parametric multiplication is performed multiple times.

In contrast, the exponentiation method based on Formula (5) achieves nearly the same performance as the classical algorithm.

This demonstrates that the use of a closed-form expression is effective for optimizing parametric exponentiation operations.

This behavior can be explained by the fact that the closed-form formulation eliminates repeated parametric multiplications, thereby reducing computational overhead and improving scalability.

The exact numerical values corresponding to the results shown in Fig. 2 are presented in Table III.

The table III shows that the algorithm based on Formula (3) is slower than the classical method, whereas the one based on Formula (5) has nearly the same speed.

TABLE III. EXECUTION TIME OF ALGORITHMS FOR DIFFERENT MODULUS SIZES

n (bit)	Classic (ms)	Param5 (ms)	Param7 (ms)
512	1.01	4.21	0.74
1024	2.64	9.32	2.66
2048	9.85	22.49	9.96
4096	37.76	63.66	38.16
5120	57.57	90.16	58.25

In addition to runtime measurements, memory usage and CPU ticks were also evaluated for different modulus sizes. The corresponding results are presented in Table IV.

TABLE IV. MEMORY USAGE AND CPU TICKS FOR DIFFERENT MODULUS SIZES

n (bit)	Classic		Param5		Param7	
	Memory(bytes)	CPU Ticks	Memory(bytes)	CPU Ticks	Memory(bytes)	CPU Ticks
512	112	21240	112	35332	112	6300
1024	112	21152	112	70439	112	21387
2048	112	74823	0	162860	112	75852
4096	0	283699	0	464917	112	287081
5120	112	446870	0	665616	112	442620

As shown in Table IV, the iterative method based on Formula (3) consistently requires more CPU ticks than the classical and closed-form methods across all tested modulus sizes. In contrast, the method based on Formula (5) remains very close to the classical algorithm, while memory usage stays nearly constant for all approaches.

The obtained results further confirm that the computational overhead associated with iterative parametric exponentiation significantly increases execution time and CPU ticks, whereas the closed-form approach mitigates this effect while preserving low memory usage.

B. Impact of the Parameter R

The performance of the algorithms was evaluated for different bit-lengths of the parameter R. This allows assessment of the effect of parameter size on execution time, memory usage, and CPU ticks. The values used in the computations were fixed as $n = 4096$bits, $a = 512$bits, and $k = 256$bits.

The bit-length of parameter R was gradually increased starting from 128 bits, and for each value, the execution time of the algorithms was measured. The obtained results are presented graphically in Fig. 3.

As the graph shows, the execution time of the classical exponentiation algorithm remains unchanged since it is independent of the parameter R. By contrast, the computation time of the exponentiation algorithm based on Formula (3) increases as the value of R grows.

Meanwhile, the parametric exponentiation algorithm based on Formula (5) performs similarly to the classical algorithm.

This indicates that the proposed method is largely independent of the parameter R, as the computation is reduced to a single modular exponentiation that does not involve repeated parameter-dependent operations.

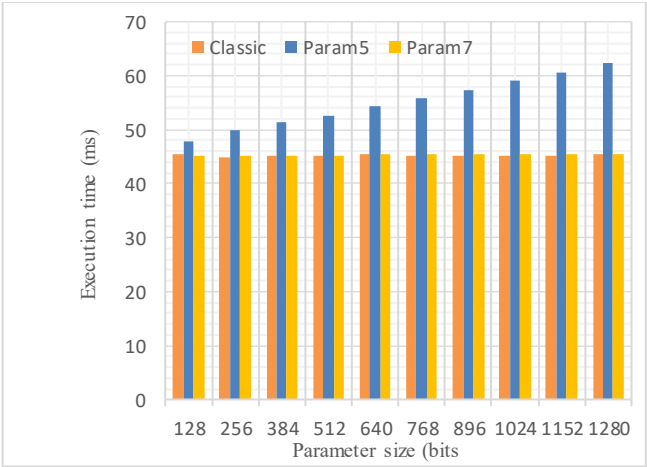


Fig. 3. Execution time vs. parameter length.

The exact numerical values corresponding to the results shown in Fig. 3 are presented in Table V.

TABLE V. EXECUTION TIME OF ALGORITHMS FOR DIFFERENT PARAMETER VALUES.

R (bit)	Classic (ms)	Param5 (ms)	Param7 (ms)
128	38.536	40.872	38.678
256	38.595	42.447	38.563
512	38.373	45.497	38.609
1024	38.386	52.356	38.721
1280	38.424	55.575	38.711

As can be seen from the table, the execution time of the algorithm based on Formula (3) increases with the growth of the parameter R. In contrast, the algorithm based on Formula (5) is almost unaffected by changes in the parameter value.

TABLE VI. MEMORY USAGE AND CPU TICKS FOR DIFFERENT PARAMETER VALUES

R (bit)	Classic		Param5		Param7	
	Memory(bytes)	CPU Ticks	Memory(bytes)	CPU Ticks	Memory(bytes)	CPU Ticks
128	112	295771	0	314064	112	295716
256	0	294361	0	326755	112	296239
512	0	295706	0	377922	112	298236
1024	0	309788	0	429285	112	308696
1280	0	309116	0	432376	112	304334

Some memory measurements are reported as zero because the benchmark records only additional managed heap allocations. In several runs, no extra allocation was detected during execution.

As shown in Table VI, CPU tick measurements clearly indicate that the iterative method based on Formula (3) becomes increasingly expensive as the parameter size grows. In contrast, the method based on Formula (5) remains close to the classical algorithm across all tested parameter sizes.

Overall, the results for different parameter sizes show that increasing the bit-length of R has a significant impact on the iterative method based on Formula (3), leading to higher computational overhead. In contrast, the closed-form method based on Formula (5) remains close to the classical algorithm in both runtime and CPU ticks, demonstrating better scalability with respect to the parameter size.

VII. CONCLUSION AND FUTURE WORKS

In this study, parametric algebraic operations were systematically studied, and an algorithm for parametric exponentiation based on modular binary exponentiation was proposed. By exploiting the properties of parametric algebra, a fast exponentiation expression for parametric exponentiation was derived as:

$$a^{\backslash k} = ((1 + Ra)^k - 1) \cdot R^{-1} \, mod \, n$$

As a result, the proposed method significantly reduces the computational impact of the parameter R on parametric exponentiation while improving overall practical efficiency.

According to the experimental results, for modulus bit-lengths ranging from 512 to 5120 bits, the execution time of the algorithm based on Formula (5) is nearly identical to that of the classical exponentiation algorithm, with the difference not exceeding 1–2% on average.

At the same time, it was observed that the algorithm based on Formula (5) operates approximately 1.7–1.8 times faster than the optimized method based on Formula (3).

Furthermore, during the experiments, it was determined that increasing the bit-length of the parameter R (from 128 bits to 1280 bits) does not significantly affect the performance of the proposed algorithm based on Formula (5).

These results confirm that parametric exponentiation can be reduced to a single modular exponentiation while preserving nearly the same performance as the classical method.

The proposed approach provides a practical and efficient alternative to classical exponentiation methods, making it suitable for high-performance cryptographic applications.

Furthermore, the proposed approach achieves a favorable balance between computational efficiency, low memory consumption, and reduced processor workload, making it suitable for practical cryptographic applications.

The computational Formula (5) proposed in this study enables efficient exponentiation in encryption algorithms and standards such as RSA and ElGamal when enhanced with parametric algebraic operations.

Limitation. This study has several limitations. Although the proposed method significantly improves the efficiency of parametric exponentiation, the practical deployment of parametric cryptosystems still requires broader optimization of all parameter-dependent operations.

Existing hardware and software acceleration mechanisms are mainly designed for classical arithmetic, while specialized implementations for parametric algebra remain limited.

In addition, the correct selection of the parameter R is important because the algebraic properties of the system may depend on its value. For example, the existence of inverse elements, the applicability of the closed-form reduction in Formula (5), and the requirement to satisfy conditions such as $GCD(R, n) = 1$ may introduce additional computational overhead through GCD checks and parameter validation procedures. If these conditions are not satisfied, some algebraic operations may become invalid or require alternative computational methods.

Future work includes the implementation of the proposed method in real-world cryptographic systems and its optimization for parallel and distributed computing environments. In addition, a formal analysis of the security implications of the proposed computational reduction remains an important research direction.

Overall, the results confirm both the theoretical soundness and practical efficiency of the proposed parametric exponentiation method.

REFERENCES

[1] A. J. Menezes, P. C. Van Oorschot, and S. A. Vanstone, *Handbook of applied cryptography*. Boca Raton: CRC Press, 2018.

[2] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, 1978.

[3] X. P. Xasanov, *Takomillashgan diamatristalar algebralari va parametrli algebra asosida kriptotizimlar yaratish usullari va algoritmlari*. Toshkent: Fan, 2008. [in Uzbek]

[4] O. Ahmedova, U. Mardiyev, and O. Tursunov, "Generation and distribution of secret encryption keys with a parameter," in *Proc. International Conference on Information Science and Communications Technologies (ICISCT)*, 2020, pp. 1–4.

[5] UNICON.UZ, "Information technology. Cryptographic protection of information. Data encryption algorithm," UzNST 270:2024, 2024.

[6] UNICON.UZ, "Information technology. Cryptographic protection of information. Hash function," UzNST 285:2024, 2024.

[7] UNICON.UZ, "Information technology. Cryptographic protection of information. Processes for generating and verifying electronic digital signatures," UzNST 286:2024, 2024.

[8] UNICON.UZ, "Information technology. Cryptographic protection of information. Processes for generating and verifying electronic digital signatures based on elliptic curves," UzNST 2826:2014.

[9] O. Ahmedova, U. Mardiyev, and O. Tursunov, "Algebraic structure of parametric elliptic curves," in *Proc. International Conference on Information Science and Communications Technologies (ICISCT)*, 2021, pp. 1–3.

[10] Z. Xudoykulov and U. Xudoynazarov, "Improving the RSA public-key encryption algorithm based on parametric algebra," The Latest News and Research in Education, vol. 2, no. 12, pp. 114–122, 2025.

[11] U. Xudoynazarov, "Parametrli algebraga asoslangan el-gamal shifrlash algoritmlarini gomomorfik xususiyatini tadqiq etish," *Al-Farg'oniy avlodlari*, vol. 1, no. 4, pp. 153–157, 2023. [in Uzbek]

[12] U. Mardiyev, "Using the R parameter in a Massey–Omura protocol on elliptic curves," in *Proc. International Conference on Information Science and Communications Technologies (ICISCT)*, 2021, pp. 1–3.

[13] M. Fartitchou, H. El Marraki, L. Lafkir, A. Azzouz, K. El Makkaoui, and Z. El Allali, "Public-key cryptography behind blockchain security," in Proc. 2022 5th International Conference on Networking, Information Systems and Security (NISS), 2022, pp. 1–5.

[14] M. Gerstenhaber, "On the deformation of rings and algebras," *Annals of mathematics*, vol. 79, no. 1, pp. 59–103, 1964.

[15] A. Aggarwal and M. Isaacs, "An elementary method for fast modular exponentiation with factored modulus," arXiv preprint arXiv:2401.10497, 2024.

[16] K. Sukhorukova, "Spaces of non-additive measures generated by triangular norms," Mat. Stud., vol. 59, no. 2, pp. 215–224, Jun. 2023.

[17] A. N. Grishkov and A. V. Zavarnitsine, "Groups with triality," *Journal of algebra and its applications*, vol. 5, no. 4, pp. 441–463, 2006.

[18] J. Hall, *Moufang loops and groups with triality are essentially the same thing*. Providence: American Mathematical Society, 2019.

[19] K. El Makkaoui, Y. Lamriji, A. Beni-Hssane, Y. Maleh, and I. Ouahbi, "An overview of fast variants of the RSA cryptosystem for modern cryptography applications," EDPACS, vol. 67, no. 6, pp. 25–34, 2023.

[20] K. El Makkaoui, Y. Lamriji, I. Ouahbi, O. Nabil, A. Bouzahra, and A. Beni-Hssane, "Fast modular exponentiation methods for public-key cryptography," in Proc. 2022 5th International Conference on Advanced Communication Technologies and Networking, Dec. 2022.

[21] O. Abramkina, M. Yakubova, T. Serikov, Y. Begimbayeva, and B. Yakubov, "Implementation of lattice theory into the TLS to ensure secure traffic transmission in IP networks based on IP PBX Asterisk," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 15, no. 10, 2024.

[22] D. Swetha and S. K. Mohiddin, "Advancing quantum cryptography algorithms for secure data storage and processing in cloud computing," *International journal of advanced computer science and applications*, vol. 15, no. 9, 2024.

[23] [23] N. Doukas, "Fast exponentiation in Galois fields GF(2^n)," International Journal of Computing, vol. 24, no. 1, pp. 81–91, 2025. https://doi.org/10.47839/ijc.24.1.3879

[24] I. Marouf, M. M. Asad, and Q. A. Al-Haija, "Comparative study of efficient modular exponentiation algorithms," Compusoft, vol. 6, no. 8, p. 2381, 2017.

[25] K. Somsuk, N. Panit, and R. Phromloungsri, "An improved algorithm for accelerating modular exponentiation with periodic bit pattern exponents," KSII Transactions on Internet and Information Systems, vol. 19, no. 4, pp. 1267–1285, 2025.

[26] K. Zarifjon and K. Umidjon, "Symmetric encryption algorithm based on a one-way function extended by parametric algebra," Stanford Database Library of American Journal of Applied Science and Technology, vol. 5, no. 12, pp. 31–36, 2025.

An Interoperable Multi-Agent Architecture for Personalized Smart Learning Using Generative AI and Learning Analytics

Al Mahdi Khaddar, Youssef Said, Amine Dehbi, Tarik Chafiq
Laboratory of Information Processing (LTI)-Faculty of Sciences Ben M'Sick,
University Hassan II of Casablanca, Casablanca, Morocco

***Abstract*—Learning Management Systems (LMSs) remain central to digital education, but they still provide limited support for adaptive and personalized learning across heterogeneous platforms. This study proposes an interoperable smart learning architecture that integrates a Multi-Agent System (MAS), generative artificial intelligence, and learning analytics to support context-aware interventions while preserving LMS independence. Methodologically, the work follows a design-oriented research approach based on architectural modeling and scenario-based validation. The proposed framework combines Learning Tools Interoperability (LTI) 1.3, the Experience API (xAPI), Sharable Content Object Reference Model (SCORM), a Learning Record Store (LRS), and an asynchronous Extensible Messaging and Presence Protocol (XMPP)/JavaScript Object Notation (JSON) communication bus to connect intelligent services with existing LMS environments. The architecture includes tutor, assessment, recommendation, monitoring, collaboration, and profile agents coordinated through a microservices-based design. Its functional coherence is illustrated through four representative scenarios covering dropout-risk detection, targeted remediation, teacher dashboards with grade return, and collaborative feedback. The main contribution is a modular and standards-based architecture that connects analytics, agent-based orchestration, and generative AI within a closed-loop adaptation process for scalable smart learning environments.**

***Keywords*—Smart learning; multi-agent systems; generative AI; learning analytics; LMS interoperability; adaptive learning**

I. Introduction

Learning Management Systems (LMSs) such as Moodle and Canvas play a central role in digital education by supporting course delivery, communication, and assessment. However, despite their widespread adoption, most LMS platforms still provide limited support for adaptive and personalized learning at scale [1], [2]. In many cases, personalization remains restricted to static configurations, rule-based options, or isolated plugins, while intelligent services are not deeply integrated into pedagogical workflows. As higher education increasingly seeks learner-centered and data-driven environments, this limitation becomes more visible and pedagogically significant [3], [4].

At the same time, learning analytics has created important opportunities for capturing and interpreting learner behavior through digital traces. Standards such as the Experience API (xAPI) and Learning Record Stores (LRSs) enable the collection, storage, and reuse of fine-grained learning events across multiple tools and platforms [5], [6]. These technologies support monitoring, prediction, and feedback processes that can improve educational decision-making [7]. Nevertheless, in many existing systems, analytics outputs remain mainly descriptive and are not systematically transformed into timely pedagogical interventions inside the LMS [8], [9]. As a result, the adaptive loop between observation, interpretation, and educational action often remains incomplete [8].

At the same time, the use of generative artificial intelligence in education to deliver adaptive explanations, automated feedback, and personalized tutoring is gaining more attention. Several recent papers have discussed the benefits that can be offered by large language models and other artificial intelligence algorithms in improving learning processes and student engagement [10], [11], [12]. Nonetheless, many existing systems continue considering generative AI as merely another component of a conversation without embedding it into the overall educational framework along with the learning experience data, pedagogical knowledge, and limitations [13], [14]. At the same time, several challenges exist in terms of hallucination, inconsistency, trust, and pedagogical control [15], [12].

Another promising path toward intelligent learning systems would be Multi-Agent Systems (MAS). These have modularity, autonomy, and decentralized control. Agent-based technologies are applicable to many tasks including tutoring, surveillance, recommendation, evaluation, and collaborative learning [16], [17], [3]. In more recent years, microservices architecture has emerged as another promising approach to loosely coupled, scalable, and independently evolving intelligent systems [18], [19]. Many of these proposed models for education, however, are largely conceptual, platform-dependent, or disconnected from existing interoperability standards. Therefore, there remains a gap in architectural models that can integrate intelligent agents, learning analytics, artificial generative intelligence, and interoperable learning management systems [20], [13].

This fragmented landscape reveals a clear research gap. Existing works often address LMS integration, learning analytics, intelligent tutoring, agent-based adaptation, or generative AI separately, but fewer studies propose a standards-based architecture that connects these components in a closed-loop process for adaptive learning [21], [22], [7], [13]. Moreover, the literature still provides limited guidance on how such an architecture can remain modular, interoperable, and

portable across heterogeneous LMS environments while supporting both learner-facing and teacher-facing interventions. This gap is particularly important in smart learning research, where the challenge is no longer only to collect educational data or deploy isolated AI services, but to orchestrate them coherently within real institutional ecosystems [23], [24].

To address this gap, this paper proposes an interoperable smart learning architecture based on a hybrid Multi-Agent System and a service-oriented microservices design. The architecture integrates Learning Tools Interoperability (LTI) 1.3 for LMS access and service delivery, xAPI and Learning Record Stores for trace collection and analytics, and asynchronous agent coordination through Extensible Messaging and Presence Protocol (XMPP) and JavaScript Object Notation (JSON). The proposed framework aims to support adaptive explanations, remediation, recommendation, monitoring, teacher dashboards, and collaborative feedback without requiring modifications to the LMS core. In this way, the LMS is preserved as a stable host environment, while intelligent services are externalized through interoperable mechanisms [6], [13].

The objectives of this study are fourfold. First, it aims to define a modular architecture that decouples intelligent services from the LMS while preserving interoperability with existing educational platforms. Second, it seeks to integrate learning analytics, intelligent agents, and generative AI into a unified closed-loop adaptation process. Third, it aims to show how standardized traces can support both individual and collaborative pedagogical interventions. Fourth, it intends to validate the internal coherence and functional completeness of the proposed architecture through representative educational scenarios.

This work adopts a design-oriented research methodology because the main goal is not to measure learner performance through a field experiment at this stage, but to design, formalize, and theoretically validate an architectural framework for smart learning. Such an approach is appropriate for complex adaptive educational systems in which interoperability, modularity, orchestration, and deployment logic must first be specified before large-scale implementation and empirical testing [25], [26], [27], [17]. Scenario-based validation is therefore used to examine whether the proposed architecture can support complete end-to-end workflows across representative educational situations, including learner support, assessment remediation, instructor analytics, and collaborative learning.

The expected contributions of this study are both scientific and practical. From a scientific perspective, the paper contributes a unified architectural model that bridges research streams that are often treated separately, namely LMS interoperability, learning analytics, intelligent agents, microservices, and generative AI in education. From a practical perspective, the proposed framework offers institutions a reusable and extensible foundation for deploying adaptive services across heterogeneous LMS ecosystems while maintaining portability, governance, and pedagogical consistency. By organizing these components within a closed-loop architecture, the study also contributes to the design of more scalable and institutionally deployable smart learning environments [23], [24].

The remainder of the paper is organized as follows. Section II reviews the related literature and identifies the main limitations of existing work. Section III presents the research methodology. Section IV describes the proposed architecture and its main layers. Section V reports the scenario-based validation and discusses the main contributions and limitations of the framework. Finally, Section VI concludes the paper and outlines directions for future work.

II. Literature Review

Recent research has confirmed the growing importance of adaptive and personalized learning in higher education. Personalized learning environments aim to adjust content, pacing, and support according to learner characteristics, performance, and engagement. In this context, LMS platforms remain essential for course management and digital delivery, but several studies report that most existing LMSs still provide limited support for deep and dynamic personalization [4], [21]. In many cases, adaptive behavior is not embedded in the LMS architecture itself, but added through external tools or isolated modules, which limits coherence and scalability [28], [22].

Learning analytics has become one of the most promising fields of research related to the use of digital traces for better comprehension and optimization of the learning process. Various studies demonstrate the potential of learner activity data for monitoring purposes and predictive analysis [5], [7]. At the same time, interoperability standards like xAPI and cmi5 have made collecting standard learning traces across diverse learning environments and storing them in Learning Record Stores more feasible [29], [6]. However, there are also significant limitations identified in past research. For example, many systems operate on a purely descriptive level, offering dashboard tools rather than advanced decision support, while the link between the results obtained and immediate pedagogical actions is poor [30], [8].

At the same time, artificial intelligence has gained increasing attention in smart learning systems. Recent reviews emphasize the potential of AI techniques, including machine learning, deep learning, and generative AI, to support adaptive feedback, learner modeling, performance prediction, and content personalization [10], [20]. More specifically, large language models and generative AI tools have shown promise for generating explanations, recommendations, and contextualized tutoring support [11], [12]. Nevertheless, the literature also underlines major concerns related to hallucination, factual inconsistency, bias, pedagogical reliability, and ethical use in educational settings [12], [14], [15]. Another limitation is that many AI-based educational tools are developed as standalone assistants rather than being embedded into an interoperable and institutionally deployable smart learning architecture [13].

Multi-Agent Systems constitute another important research direction for intelligent learning environments. MAS-based approaches are valued for their modularity, autonomy, and distributed problem-solving capacity, making them suitable for complex educational functions such as tutoring, assessment, recommendation, and monitoring [3], [16]. In parallel, microservice-based architectures have been recognized as effective for building scalable, loosely coupled, and evolvable intelligent systems [18], [19]. These properties are particularly

relevant for educational ecosystems that need to integrate multiple intelligent services and continuously evolving AI models. However, several previous works remain either conceptual, prototype-level, or narrowly focused on one pedagogical function. In addition, many of them do not explicitly support standards-based interoperability with mainstream LMS platforms, which reduces portability and institutional applicability.

Another limitation in the existing literature is the fragmented treatment of the main components of smart learning. Some studies focus on LMS-centered personalization, others on learning analytics pipelines, others on generative AI assistants, and others on agent-based educational systems [31], [7], [13], [21]. While each stream offers useful contributions, fewer works integrate these components into a unified architecture that supports trace collection, feature computation, learner modeling, agent coordination, and delivery of interventions back into the LMS through interoperable standards. This gap becomes even more significant when both learner-facing and teacher-facing interventions are considered within the same framework.

To clarify the position of the present study, Table I compares representative previous works with the proposed approach according to key dimensions: architectural scope, interoperability support, analytics integration, agent-based intelligence, generative AI integration, and validation approach. The comparison shows that previous studies generally address only part of the problem. Some works emphasize analytics but do not include agent orchestration. Others integrate AI assistants into LMSs but do not provide a broader closed-loop adaptation model. Similarly, some studies propose intelligent educational services, yet without strong interoperability support or without considering portability across heterogeneous LMS environments [11], [13], [29], [31].

Based on this analysis, the novelty of the present work lies not in introducing a single isolated technology, but in integrating several complementary dimensions into one coherent architecture. The proposed framework combines standards-based LMS interoperability, xAPI-based trace collection, Learning Record Store-centered analytics, a hybrid Multi-Agent System implemented as microservices, and generative AI-enabled pedagogical services within a closed-loop adaptation process. In addition, the architecture is designed to support multiple educational situations, including dropout-risk detection, targeted remediation, teacher dashboards, and collaborative feedback. This combination distinguishes the proposed work from prior approaches that are typically narrower in scope, less interoperable, or insufficiently connected across analytics, intelligence, and pedagogical action [23], [24].

TABLE I. COMPARISON OF RELATED WORKS AND THE PROPOSED APPROACH

Ref.	Main Focus	LMS Interop.	Learn. Analytics	Multi Agent Supp.	Gen AI Integ.	Main Lim.
[4]	Personalized adaptive learning in HE	No explicit standards focus	Partial	No	No	Focuses on adaptive learning characteristics rather than system architecture
[7]	ML and GenAI in learning analytics	Partial	Yes	No	Yes	Reviews models and trends, but does not propose deployable architecture
[29]	Predictive analytics for course completion	Partial	Yes	No	No	Strong analytics pipeline, but no agent orchestration or LMS-wide adaptation model
[13]	GenAI assistants integrated into LMS	Yes	Partial	No	Yes	Focuses on LMS assistant integration, not full multi-agent closed-loop architecture
[11]	AI and analytics for pedagogical decisions	Partial	Yes	No	Yes	Strong decision support perspective, but limited architectural interoperability detail
[30]	Adaptive GenAI-based feedback and recommendation	No explicit standards focus	Yes	No	Yes	Focused on a specific learning system, not heterogeneous LMS environments
Proposed work	Interoperable smart learning architecture	Yes	Yes	Yes	Yes	Scenario-based validation only; empirical deployment remains future work

III. MATERIALS AND METHODS

This study adopts a design-oriented research methodology focused on the conception, formalization, and theoretical validation of an interoperable smart learning architecture. The purpose of the work is not to evaluate learner outcomes through a classroom experiment at this stage, but to develop a coherent architectural framework capable of integrating intelligent agents, learning analytics, generative artificial intelligence, and Learning Management System (LMS) interoperability within a unified smart learning ecosystem. This methodological choice is appropriate because the contribution of the paper is primarily architectural and system-oriented rather than experimental. It also responds to the need for a structured design process before empirical implementation in authentic educational environments [25].

The methodological process began with the identification of the research problem through an analysis of limitations reported in previous studies on LMS-centered personalization, learning analytics, multi-agent educational systems, and generative artificial intelligence in education. This analysis revealed a fragmented landscape in which these components are often treated separately. In particular, the literature showed the absence of a unified framework capable of combining interoperability, modularity, analytics, and intelligent pedagogical intervention within a closed-loop process. Based on

this gap, the main functional and technical requirements of the proposed architecture were defined, including LMS independence, standards-based interoperability, modularity, scalability, portability across heterogeneous platforms, and support for adaptive learner-facing and teacher-facing interventions [7], [13], [21], [22], [31].

Fig. 1 presents the overall methodological workflow followed in this study, from problem identification and requirement specification to layered architecture design and scenario-based validation.

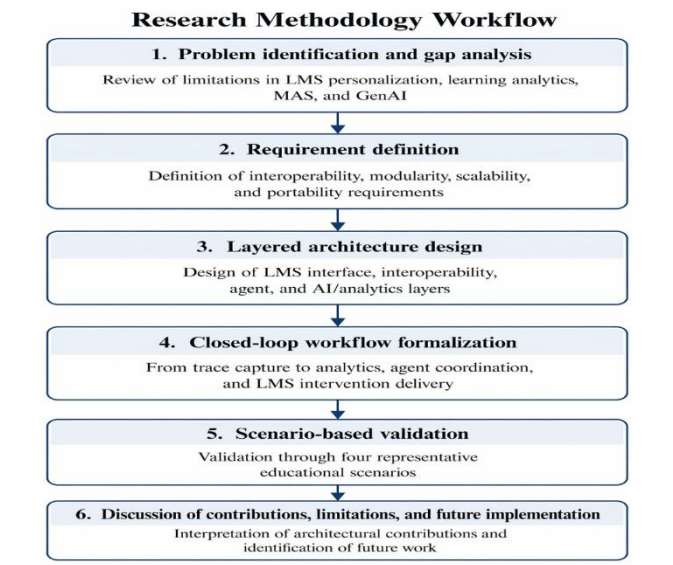


Fig. 1. Research methodology workflow.

The proposed architecture was then designed using a layered model composed of four main layers: the LMS interface layer, the interoperability layer, the intelligent agent layer, and the artificial intelligence and learning analytics layer. This decomposition was adopted to reduce coupling between components and to facilitate extensibility, maintainability, and cross-platform portability. In the proposed design, the LMS acts as a host environment that provides access to courses, users, and activities, while intelligent services are externalized and connected through interoperability standards. This architectural choice makes it possible to preserve the stability of the LMS core while enabling the integration of adaptive services developed independently of the platform. This point is particularly important because the practical value of the proposed work depends on its ability to operate across heterogeneous LMS environments without requiring changes to internal LMS code [18], [19], [27].

To specify the internal operation of the framework, the architecture was formalized as an end-to-end closed-loop workflow. Learner and teacher interactions are captured through interoperable mechanisms, represented as standardized events, and stored in a Learning Record Store. These traces are then processed by the analytics layer to compute indicators related to engagement, progression, difficulty, or risk, and to update the learner model or feature store. The resulting features are consumed by intelligent agents, which coordinate through asynchronous communication in order to produce pedagogical actions such as feedback, recommendations, remediation, alerts, or dashboard content. These actions are finally delivered back to the LMS through interoperable services. This formalization ensures that the proposed framework is not limited to a static architectural description, but includes the logical sequence of adaptive processing steps from data capture to pedagogical intervention, as requested by the reviewers [16].

At the technical level, the architecture relies on several complementary technologies and standards. Learning Tools Interoperability 1.3 is used for secure tool launch, contextual access, and LMS-facing service delivery. The Experience API is used to represent learning interactions as standardized statements, while a Learning Record Store functions as the central repository for trace collection. Intelligent services are modeled as distributed microservices, and agent coordination is supported through Extensible Messaging and Presence Protocol and JavaScript Object Notation message exchange. In addition, the analytics layer includes a feature computation process combining rule-based indicators and optional machine learning inference, depending on the educational scenario and the availability of predictive models. This technical description clarifies the implementation logic of the framework and strengthens the architectural consistency of the proposed model [6], [15], [18], [19].

The theoretical validation of the proposed architecture was conducted through a scenario-based approach. Four representative educational scenarios were defined to examine whether the framework can support complete workflows across different pedagogical situations: early dropout-risk detection, targeted remediation after quiz failure, teacher dashboard generation with grade return, and collaborative group feedback. These scenarios were selected because they cover different categories of educational intervention, namely individual monitoring, assessment support, teacher-facing analytics, and collaborative learning support. The aim of this validation is not to measure statistical performance, latency, or scalability, but to assess the functional completeness, internal coherence, and standards-based interoperability of the proposed framework across representative use cases. This clarification is important because the study is positioned as an architectural contribution and not as an empirical performance evaluation [26], [27].

Although this study does not yet include an empirical deployment in an authentic classroom setting, the adopted methodology provides a rigorous foundation for future implementation. By explicitly defining the architecture, its components, their interactions, and the logic of validation, the study establishes a reproducible design framework that can later be extended toward prototyping, simulation, and real-world experimentation. In this sense, the methodological contribution of the paper lies in transforming a fragmented set of technological and pedagogical components into a structured, interoperable, and theoretically validated smart learning architecture [25], [26], [27].

IV. Proposed Approach

This paper proposes an interoperable intelligence layer that operates independently of the Learning Management System (LMS) and enables the deployment of adaptive smart learning services across heterogeneous educational platforms. The main

idea is to treat the LMS as a host environment responsible for course access, user context, and activity management, while intelligent services are externalized as interoperable components connected through standard protocols. This design allows institutions to integrate, replace, or extend intelligent educational services without modifying the LMS core, thereby improving portability, modularity, and long-term maintainability [13], [18], [19].

The proposed approach is guided by four design principles. The first principle is interoperability-first integration, which ensures that intelligent services interact with existing educational platforms through standards-based mechanisms rather than proprietary interfaces. The second principle is modularity through distributed agent microservices, which makes it possible to isolate pedagogical functions and evolve them independently. The third principle is data-driven adaptation based on standardized learning traces, allowing pedagogical interventions to be grounded in learner and group behavior. The fourth principle is governance by design, meaning that identity, access control, and privacy constraints are embedded in the architecture from the outset [6], [18].

The proposed architecture is structured into four interconnected layers: the LMS interface layer, the interoperability layer, the intelligent agent layer, and the artificial intelligence and learning analytics layer. This layered decomposition provides a clear separation of responsibilities while preserving end-to-end coordination across the system. It also enables the architecture to remain LMS-agnostic, since adaptive services are not embedded directly in LMS logic but operate through interoperable connections [13], [18], [19]. Fig. 2 illustrates an interoperable smart learning architecture based on a hybrid multi-agent system and service-oriented architecture integrating intelligent agents, generative AI, and learning analytics.

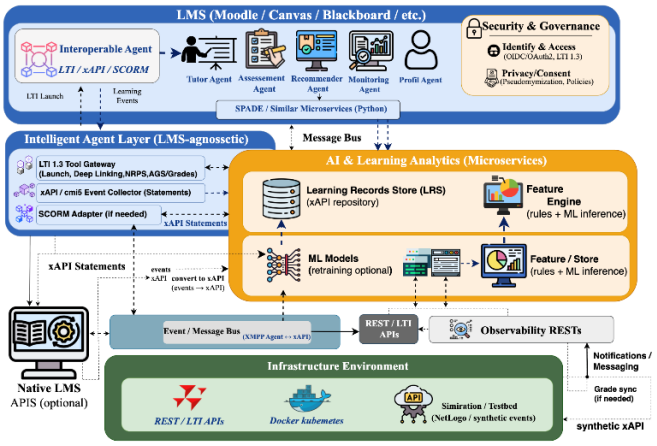


Fig. 2. Interoperable smart learning architecture based on a hybrid Multi-Agent System and Service-Oriented Architecture integrating intelligent agents, generative AI, and learning analytics.

A. Layered Architecture

The architecture is structured into four layers: 1) LMS interface, 2) interoperability, 3) intelligent agents, and 4) AI and learning analytics, deployed on a containerized infrastructure.

1) LMS interface layer: At the LMS interface layer, platforms such as Moodle, Canvas, or Blackboard remain the primary environments in which learners and instructors access courses, assessments, and collaborative activities. In the proposed model, the LMS is not redesigned at the core level. Instead, it functions as a stable delivery environment in which adaptive services are made available through external interoperable tools. This choice preserves institutional compatibility with existing LMS infrastructures while allowing intelligent functionalities to evolve outside the platform [1], [2], [13].

2) Interoperability Layer (LTI / XAPI / Cmi5 / SCORM): The interoperability layer facilitates the communication between the LMS and external intelligence services. The interoperability layer mainly uses Learning Tools Interoperability (LTI) 1.3 as the means of launching applications securely, accessing contextual data, and integrating services within the LMS environment. External tools will have access to contextual data, including the course identifier, the user's role, and activity contexts. Moreover, external tools may send back results, including grading and embedded pedagogic services, where required. Simultaneously, the Experience API (xAPI) collects learner interactions as structured statements, with CMI5 profiles available for cases where there is a need for learning activities to be packaged into structures. The learning statements are recorded in the Learning Record Store (LRS) module, enabling a consolidation of learning tracks from both the LMS and external tools. Finally, there is an optional SCORM adapter that can be managed to support legacy learning content [5], [6], [29].

3) Intelligent agent layer: At the core of the architecture, the intelligent agent layer provides the main adaptive and decision-making functions of the system. This layer is implemented as a distributed set of microservices, with each agent responsible for a specific pedagogical role. The tutor agent supports adaptive explanations, guidance, hints, and remediation messages. The assessment agent manages quiz-related diagnosis, feedback generation, and grading-oriented pedagogical logic. The recommender agent selects suitable resources, activities, or follow-up actions according to learner state and contextual needs. The monitoring agent tracks engagement, inactivity, progression, and risk indicators in order to trigger timely interventions. The collaboration agent analyzes group activity and supports collaborative learning regulation. The profile agent maintains the learner model by managing persistent features and learner descriptors used for personalization [16], [17], [18], [19].

To strengthen the technical soundness of the architecture, the decision logic of the agents is based on three complementary mechanisms. First, rule-based logic is used for explicit pedagogical conditions, such as prolonged inactivity, low quiz scores, missing milestones, or unbalanced group contribution. Second, analytics-derived features extracted from learner traces inform the interpretation of learner state, including engagement, progression, mastery, and risk indicators. Third, optional

machine learning inference can be incorporated when predictive models are available, for example, to estimate dropout risk, classify engagement patterns, or rank recommended resources. In this way, agents do not operate as opaque black boxes, but as orchestrated services that consume structured inputs and produce pedagogically meaningful actions according to explicit functional roles [5], [7], [16], [29].

Agent coordination is achieved through asynchronous communication over an Extensible Messaging and Presence Protocol (XMPP) and JavaScript Object Notation (JSON) message bus. This event-driven communication model allows the agents to remain loosely coupled while supporting multi-step workflows. For example, a Monitoring Agent may publish a risk signal, which is then consumed by the Recommender Agent to select a follow-up activity and by the Tutor Agent to formulate a guidance message. This asynchronous orchestration is important in smart learning environments because it supports distributed decision-making, flexibility, and the independent evolution of services [18], [19].

4) AI and learning analytics layer: The artificial intelligence and learning analytics layer transforms raw learning traces into actionable information. The learning record store functions as the central repository of XAPI statements produced by LMS activities and interoperable tools. A feature engine processes these traces and computes pedagogically relevant indicators such as time on task, login frequency, completion rate, mastery progression, error patterns, group participation balance, and inactivity duration. These indicators are then stored in a feature store or learner model database, which provides a longitudinal representation of learners and groups. When predictive models are available, machine learning inference can be incorporated into this layer in order to generate higher-level outputs such as risk scores, recommendation rankings, or engagement classifications [29].

Generative artificial intelligence is integrated into the proposed architecture as a pedagogical support component rather than as an isolated conversational tool. Within this framework, generative AI can support adaptive explanations, remediation messages, tutoring feedback, and contextual recommendations. However, its use is constrained by pedagogical and institutional control. The output of generative services is expected to be grounded in learner context, analytics-derived features, and predefined pedagogical objectives. In addition, because generative AI may produce hallucinations or factually inconsistent outputs, the architecture assumes the use of guardrail mechanisms such as prompt restriction, context grounding, pedagogical templates, and validation rules before delivery in sensitive educational situations [10], [11], [15].

B. End-to-End Closed-Loop Operation

A central contribution of the proposed approach lies in the end-to-end closed-loop adaptation process. Learner and teacher interactions generate standardized traces, which are stored and analyzed to compute features and update the learner model. Intelligent agents then consume these features, coordinate with one another, and determine which pedagogical intervention should be produced. Finally, the selected intervention is delivered back into the LMS through interoperable services such as LTI tools, dashboards, recommendation panels, feedback widgets, or grade return interfaces. This closed-loop operation ensures that data collection, interpretation, decision-making, and pedagogical action are connected within the same architectural framework [6].

C. Security, Privacy, and Institutional Constraints

Because the architecture is designed for institutional deployment, governance and privacy constraints are considered from the outset. Identity and access control rely on secure interoperability mechanisms associated with LTI 1.3. Trace collection and learner modeling should follow data minimization principles, and pseudonymization may be applied when possible. These measures are important for ensuring that adaptive services remain compatible with institutional policies related to security, privacy, and responsible artificial intelligence use [6], [14], [26].

Overall, the proposed approach differs from narrower educational architectures by integrating LMS interoperability, standardized trace collection, distributed intelligent agents, learning analytics, and generative artificial intelligence within one coherent framework. Its novelty lies not in the isolated use of a single technology, but in the explicit orchestration of these components in a modular and closed-loop architecture capable of supporting learner-facing, teacher-facing, and collaborative interventions across heterogeneous LMS environments [24].

D. Scenario-Based Validation

To validate the proposed closed-loop architecture, we adopt a scenario-based validation approach. Rather than providing a statistical evaluation of performance, this approach is intended to demonstrate that the architecture is functionally complete, standards-compliant, and internally consistent across representative educational situations. Each scenario is analyzed by following an end-to-end interaction chain: 1) LTI-based access to the learning tool, 2) generation and storage of xAPI traces in the Learning Record Store (LRS), 3) feature computation and learner model update, 4) agent coordination through the message bus, and 5) delivery of adaptive interventions back to the LMS via LTI.

1) Validation method and notation: for each scenario Si, the validation process is described using a common analytical structure that captures the main functional components involved in the closed-loop workflow. As summarized in Table II, this structure includes the actors, the learning context, the observed events, the derived features, the agent decisions, and the intervention actions returned to the LMS.

TABLE II. SCENARIO VALIDATION METHOD AND NOTATION

Element	Description
Actors	Learner, instructor, LMS, and external tool or intelligent agent
Context C	Course setting, activity type, assessment constraints, and user roles
Observed events E = {e1, ⋯,en}	Learning events captured as xAPI statements

Derived features F	indicators computed by the Feature Engine through rule-based processing and, where applicable, machine learning inference
Agent decisions D	Decisions produced by the intelligent agents, along with the coordination messages exchanged through the XMPP/JSON message bus (SPADE)
Intervention actions A	Adaptive actions are delivered back to the LMS through LTI 1.3, and grades are returned through standard LTI services when required.

2) Validation Criteria: For each scenario, the proposed approach is validated through six evaluation criteria, which are detailed in Table III.

TABLE III. SCENARIO-BASED VALIDATION CRITERIA

N	criteria	Definition
C1	Traceability	Each interaction provided to the learner or tutor must be traceable through an established chain, which consists of captured events, calculated indices/attributes, and the consequent decision taken by the agent.
C2	Closed-loop completeness	It is essential that the complete adaptive cycle takes place without any gaps: context capture, event capture, learning record storage, analysis, learner model modification, decision-making by agents, and the final presentation within the LMS environment.
C3	LMS Decoupling	The scenario should not involve changes to the source code or proprietary API of the LMS. It must use interoperability standards and third-party service providers while using LMS as the hosting platform.
C4	Governance consistency	Authentication, authorization, and data access should adhere to the chosen governance strategies (authentication flows for LTI 1.3, access control) and privacy requirements (data minimization, pseudonymization if possible)
C5	Agent substitutability.	A replacement for any given agent is one that adheres to the same set of contracts: the same schema for input features, the same schema for actions, and the same message structure on the bus. Everything else in the system stays the same.
C6	Cross-platform portability	This scenario will be applicable even when there is an LMS change, as long as the new LMS has support for the same interoperability endpoints. There will be no need to alter anything apart from the configuration and registration parameters on the LMS side.

a) Scenario 1: Early dropout detection and personalized reactivation in the LMS: This scenario aims to validate that the proposed architecture can detect dropout risk at an early stage from learner activity traces and then trigger an appropriate intervention directly within the LMS. The objective is not merely to observe indicators, but to demonstrate the operation of the closed-loop adaptation cycle: interaction capture, analysis, agent-based decision-making, and delivery of a pedagogical action through an interoperable tool, as illustrated in Fig. 3.

A learner is enrolled in a course on Moodle. After an initial period of normal activity, the learner's engagement decreases (fewer logins, lack of progress, and limited interaction with learning resources). The system should identify this situation as a risk signal and propose a re-engagement action.

Step 1: Context and interaction capture (Moodle + LTI 1.3 + xAPI): The learner interacts with the course in Moodle. Intelligent services are accessed through LTI 1.3 tools, which provide secure access and learning context (course, role, activity). Relevant interactions generate learning events represented as xAPI statements (typical examples include opening an activity, viewing a resource, quiz attempt/abandonment, time spent, and progression events).

Step 2: Storage and processing (LRS + Feature Engine): The xAPI statements are stored in the Learning Record Store (LRS), which serves as the unified source of learner traces. The Feature Engine processes these traces to compute engagement indicators such as session frequency (number of logins per week), time-on-task, activity completion rate, inactivity duration (days since last interaction), and progression across course modules.

These indicators then persisted in the Feature Store / Learner Model to maintain a longitudinal state of the learner.

Step 3: Decision and orchestration (Monitoring Agent + XMPP/JSON bus): The Monitoring Agent periodically retrieves indicators from the Feature Store and applies the risk detection logic (pedagogical rules and/or ML inference, depending on the chosen design).

When risk conditions are met (e.g., prolonged inactivity and low progression), the Monitoring Agent publishes a message on the XMPP/JSON message bus (SPADE). This publication may trigger multi-agent coordination:

- The Recommender Agent selects a simple next action (a priority activity or a short resource).
- The Tutor Agent prepares a guidance/remediation message (e.g., reminder + clear instruction + suggestion for gradual re-entry).

Step 4: Delivery of the intervention in Moodle (LTI tool): The intervention is displayed directly in Moodle through an LTI component integrated into the course interface (widget, recommendation panel, or notification). The re-engagement content typically includes:

- A contextualized reminder (overdue activity, module to resume).
- A concrete recommendation (e.g., a short resource or a self-assessment exercise).
- A brief tutoring message (re-entry advice, next steps).
- Optionally, a follow-up mechanism (e.g., "I resumed" confirmation, "Start now" button).

This scenario also proves that operational adaptation is possible through this architecture, where standardized traces (xAPI) are used for engagement indicators, which then trigger an agent-based decision and eventually a direct action in the LMS through the use of LTI. The stability of the LMS is also ensured, where no internal changes are needed, and the intelligence is provided through the analytics layer and agent layer, proving the principles of decoupling and portability through other standards-based LMS systems.

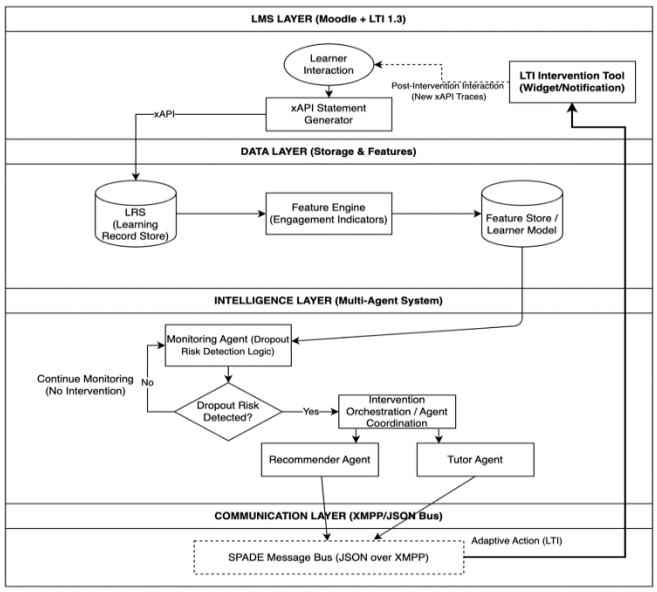


Fig. 3. Scenario 1: Early dropout risk detection and personalized re-engagement within the LMS.

b) Scenario 2: targeted remediation after quiz failure, with recommendations and tutoring in Moodle: this scenario is meant to test the fine-grained pedagogical adaptation based on assessment data. The idea is to demonstrate that assessment data, represented as standardized traces, can be used to perform a diagnostic process (e.g., poor concepts learned) and subsequently execute a specific action: providing feedback, learning materials, and tutorial support, all via lti within moodle itself, as illustrated in Fig. 4. This scenario, therefore, serves as another test of the consistency of the closed-loop process and the effectiveness of the multi-agent collaboration.

A learner is taking a quiz in Moodle or an external tool. The learner is getting a low grade or consistently answering a question incorrectly. The system must interpret the actions of the learner, determine the source of the error, and decide the best remediation.

Step 1: Capture of results and interactions (Moodle + LTI + xAPI): The learner launches the quiz, answers the questions, and either completes or abandons the attempt. Assessment-related events are captured as xAPI statements, for example: quiz start, correct/incorrect answers, final score, number of attempts, and response time.

If the quiz is associated with concepts (e.g., question tags or learning objectives), these metadata enable a more precise diagnosis.

Step 2: Storage and analysis (LRS + Feature Engine): The statements are stored in the LRS. The Feature Engine then computes indicators such as overall score, error rate by concept (when available), most problematic questions, abnormally high or low response time, and progression across attempts.

These data update the learner model/feature store, including the estimated mastery level for each concept, recurrent difficulties, and progression state.

Step 3: Decision and orchestration (Assessment Agent, Tutor Agent, Recommender Agent): An assessment agent, or the analysis logic itself, detects a remediation condition (e.g., low mastery of a concept). Then, through the XMPP/JSON bus (SPADE):

- The recommender agent selects resources adapted to the weak concept (lesson, short video, simple exercise),
- The tutor agent generates pedagogical feedback (explanation of the typical error, concept reminder, worked example),
- Optionally, the monitoring agent records that the learner is having trouble, enabling longitudinal follow-up.

Step 4: Delivery of remediation in Moodle (LTI tool): In Moodle, an LTI panel displays:

- A short explanatory feedback message (what was not understood).
- A prioritized recommendation of resources.
- A reinforcement activity (mini-exercise, short quiz, checklist).
- A suggested recovery strategy (e.g., review section X, then retry the quiz).

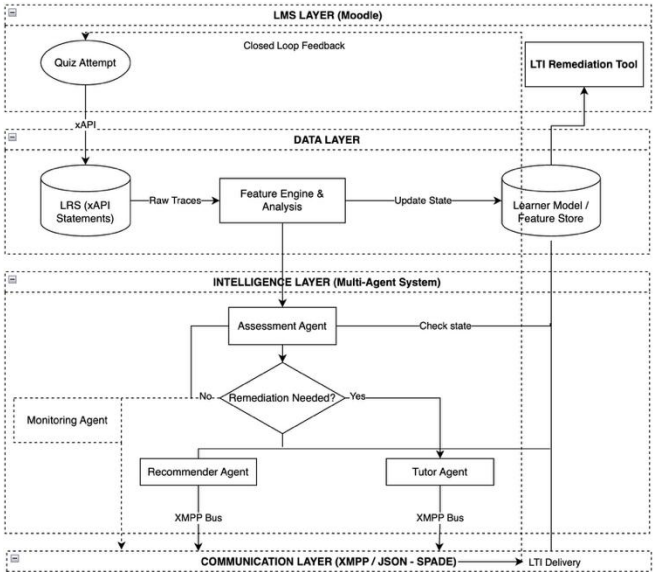


Fig. 4. Scenario 2: Targeted remediation after quiz failure, with recommendations and tutoring in Moodle.

In this scenario, architecture can take the results of assessment and translate them into active pedagogical action without the need for an LMS. This is because the same agents and analytics services could be utilized with any standards-based LMS platform due to the input/output remaining the same.

c) Scenario 3: Teacher dashboard and grade return: the above situation serves to validate the institutional and operational aspects of the proposed approach. In addition to the learner-focused services, the system needs to offer aggregated analytics that are relevant to the instructor and grade return via standard services when needed. As illustrated in Fig. 5,

architecture enables the collection, analysis, and visualization of assessment results to support pedagogical decision-making at the course level.

The assessment activity may be a quiz, an assignment, or an external learning activity. The results are individual. The teacher needs a dashboard to get an overall picture of how the class is performing. Additionally, the teacher needs a dashboard to identify areas where the class may be struggling. Finally, the teacher needs a dashboard to automatically sync the results back into the Moodle site if the assessment activity is an external tool.

Step 1: Capture of results (LTI + xAPI): The assessment can be a native one or an external one via an LTI 1.3-enabled tool. The results can be scored, successful or unsuccessful, progression, time spent, or attempts. These are then sent to the central LRS.

Step 2: Computation of cohort metrics (Feature Engine + analytics): The analytics engine computes class-level indicators, such as:

- Mean, median, and score distribution,
- Success rate by question or concept,
- Difficulty indicators,
- Detection of at-risk groups (e.g., learners showing low progression),
- Correlation between engagement and performance (when available).

These metrics are stored in the Feature Store and made available to visualization services.

Step 3: Dashboard generation (dashboard service): A dashboard service can be used to obtain the aggregated metrics, which are used to create a view for the teacher. This view is embedded within the LMS as a tool using the LTI, so the teacher can work within the familiar environment. The dashboard can display:

- Overall cohort progression.
- Problematic concepts.
- A list of learners requiring follow-up.
- Pedagogical recommendations (e.g., revisiting concept X in class).

Step 4: Grade return to Moodle (LTI services): If the assessment activity is externalized, the grades can be sent back to Moodle via the usual services provided by LTI (grade synchronization when needed). The grade is then visible in the gradebook without any manual intervention.

This scenario validates the architecture's ability to support institutional needs: cohort analytics, instructor-facing interfaces integrated into the LMS, and standards-based grade management. It strengthens the argument for deployable architecture in real educational environments, where governance, LMS stability, and interoperability are essential.

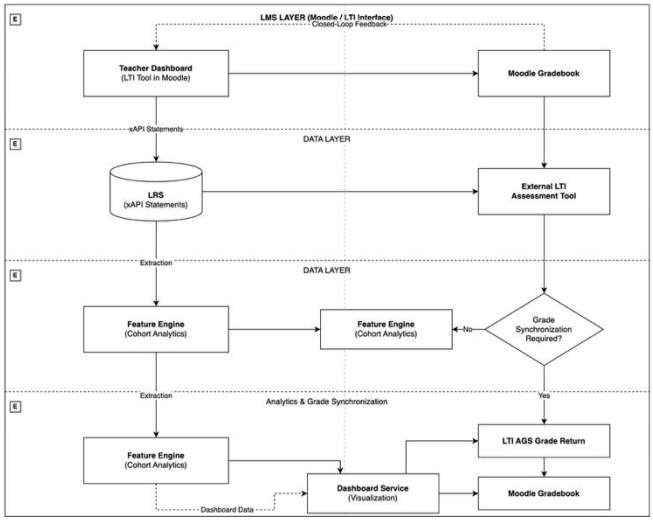


Fig. 5. Scenario 3: Teacher dashboard and grade return.

d) Scenario 4: Group interaction analysis and collaborative feedback integrated into Moodle: This scenario is intended to validate that not only is this architecture applicable for individual adaptations, but also that it is applicable for collaborative learning situations, including group work, forums, projects, etc. The purpose is to validate that traces of interaction between learners can be gathered in a standard format, analyzed to detect patterns of collaboration (such as unbalanced participation, no coordination, low levels of interaction), and then converted into pedagogical interventions that are directly provided inside Moodle through LTI tools, as illustrated in Fig. 6.

The learners are working on a collaborative activity such as a project, forum discussion, wiki, document, or integrated external tool. Some groups may be successful, while others may not be successful. This may be shown by one learner completing the entire project, low overall participation, low levels of communication among the group members, or the group failing to meet the deadlines. The system needs to recognize such groups and give them feedback.

Step 1: Capture of collaborative interactions (Moodle + tools + xAPI): Actions related to collaboration are recorded in the form of events like posts, replies, comments, file submissions, document editing, messages, validation of milestones, meetings, etc. These actions are then sent out in the form of xAPI statements. The traces may originate from:

- Native Moodle tools (forum, assignment, group activity).
- Or external tools integrated through LTI 1.3.

Step 2: Storage and analysis (LRS + Feature Engine): The LRS centralizes the statements. The Feature Engine computes individual- and group-level indicators, for example:

- Contribution distribution across group members.
- Interaction and reply frequency within the group.

- Response delays.
- Regularity of participation over time.
- Progress with respect to project milestones.
- Coordination indicators (activity concentrated on one learner, periods of collective inactivity).

These indicators feed the feature store (group profile and individual profiles), enabling longitudinal analysis.

Step 3: Decision and orchestration (Collaboration Agent + Monitoring Agent): The collaboration agent exploits group-level metrics to detect patterns such as:

- Unbalanced participation (one dominant contributor).
- Low overall activity (inactive group).
- Lack of exchanges (few replies or interactions).
- Risk of delay (slow milestone progression).

When no critical pattern is detected, the system continues monitoring the group without triggering intervention. When a relevant pattern is detected, the Collaboration Agent initiates the intervention flow through the SPADE bus (XMPP/JSON). Depending on the case, this may coordinate:

- The Monitoring Agent to generate a "group at risk" alert.
- The Tutor Agent to formulate collaborative feedback or guidance.
- Optionally, the Recommender Agent suggests resources on project management, role distribution, or coordination strategies.

Step 4: Delivery of feedback in Moodle (LTI tools): The system delivers the intervention through an LTI tool integrated into Moodle. Once the intervention has been coordinated through the SPADE bus, the Feedback Delivery Service receives the outputs produced by the agents, aggregates the relevant alerts, recommendations, and tutoring feedback, and transforms them into an LTI-compatible payload. It then packages the pedagogical action into a deliverable format and transmits it to the LTI Intervention Tool for direct presentation within Moodle. Depending on the pedagogical policy, the intervention may be:

- Addressed to the group: feedback message, coordination advice, task checklist, role distribution suggestion,
- Addressed to the teacher: group dashboard, alerts on at-risk teams, contribution summary,
- Or mixed: lightweight feedback to the group plus a detailed alert to the teacher.

This scenario also proves that architecture supports collaboration analytics/intervention without relying on a particular LMS solution. The traces are standardized through xAPI, and the delivery is done through LTI, making this mechanism easily transportable to other standards-based LMS environments. This scenario also makes the model more relevant to institutions, since the teacher can gain insight into group collaboration.

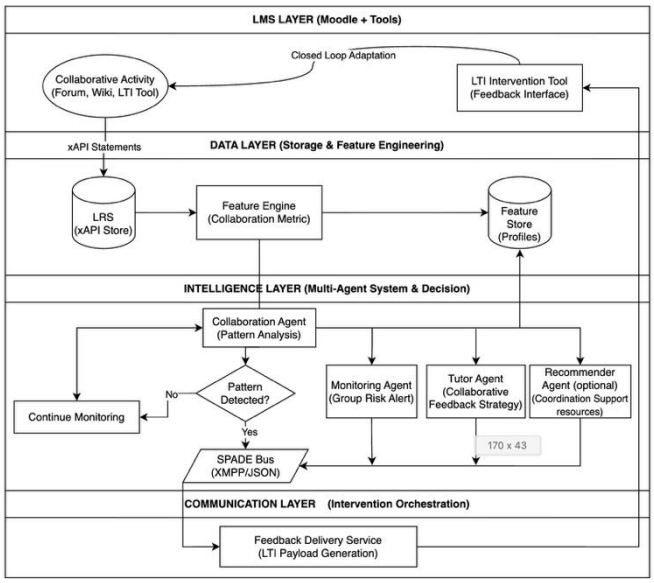


Fig. 6. Scenario 4: Group interaction analysis and collaborative feedback integrated into Moodle.

V. Result and Discussion

The results of this study are architectural and analytical rather than experimental. Accordingly, the discussion focuses on the main findings derived from the design and scenario-based validation of the proposed framework. The analysis shows that the proposed architecture contributes in four main ways: it improves interoperability across heterogeneous Learning Management System (LMS) environments, integrates analytics and intelligent agents within a closed-loop workflow, incorporates generative artificial intelligence under pedagogical and governance constraints, and supports multiple categories of educational intervention through a unified modular design.

A first important finding is that the proposed framework improves LMS decoupling and interoperability compared with many existing intelligent learning solutions. In several previous works, adaptive functions remain strongly tied to a specific platform, tool, or plugin, which limits portability and long-term maintainability. In contrast, the proposed architecture separates the LMS hosting role from the intelligence layer and relies on standards such as Learning Tools Interoperability (LTI) 1.3, the Experience API (xAPI), cmi5, and Sharable Content Object Reference Model (SCORM) to ensure communication between the LMS and external intelligent services. This design reduces platform dependency and makes it possible to deploy the same adaptive logic across multiple LMS environments without modifying the LMS core. From an institutional perspective, this is a significant advantage because it supports reuse, extensibility, and more sustainable integration of smart learning services.

A second finding is that the proposed model offers a stronger integration between learning analytics and pedagogical action than many related architectures. Several previous studies have emphasized data collection, dashboards, or predictive analytics, but fewer have connected these components to a complete adaptive intervention cycle inside the LMS [32]. In the proposed framework, learner and teacher interactions are captured as

standardized traces, stored in a Learning Record Store, transformed into pedagogically relevant indicators, and then consumed by intelligent agents that generate contextualized interventions. This closed-loop organization constitutes an important contribution because it links observation, interpretation, decision-making, and action within one coherent architecture. In this sense, the framework goes beyond descriptive analytics and moves toward operational adaptation.

A third finding concerns the role of intelligent agents in structuring adaptive decision-making. The proposed framework defines a distributed set of agents with complementary pedagogical roles, including tutoring, assessment support, recommendation, monitoring, collaboration analysis, and learner modeling. Compared with narrower agent-based educational systems, the proposed architecture does not restrict intelligence to a single function. Instead, it supports multi-agent orchestration through asynchronous communication and shared analytics outputs. The inclusion of explicit rule-based logic, analytics-derived features, and optional machine learning inference also strengthens the technical soundness of the model. This means that the intelligent agents are not only conceptual entities, but are embedded in a structured decision process that can support explainable and context-aware interventions.

A fourth finding is that the integration of generative artificial intelligence is more pedagogically controlled than in many current educational AI solutions. In numerous existing approaches, generative artificial intelligence is introduced mainly as a standalone assistant or conversational interface. In contrast, the proposed framework positions generative services within a broader architecture governed by learner context, analytics-derived features, pedagogical objectives, and institutional rules. In addition, the model explicitly acknowledges risks related to hallucination and factual inconsistency and assumes the use of guardrail mechanisms such as prompt restriction, context grounding, pedagogical templates, and validation rules. This aspect is important because it addresses a central concern raised in the recent literature and by the reviewers, namely that generative artificial intelligence should not be integrated into educational settings without mechanisms that support reliability and pedagogical control.

The scenario-based validation also highlights the versatility of the proposed architecture. Scenario 1 shows that the framework can support early dropout-risk detection and re-engagement by linking engagement indicators to personalized interventions delivered directly within the LMS. Scenario 2 demonstrates fine-grained remediation after quiz failure through the combination of assessment traces, diagnosis, tutoring feedback, and targeted recommendations. Scenario 3 extends the framework toward teacher-facing analytics and grade return, which is important for institutional deployment because smart learning systems must serve not only learners but also instructors. Scenario 4 shows that the architecture can support collaborative learning by analyzing group interactions and delivering group-oriented or teacher-oriented feedback. Taken together, these four scenarios indicate that the same architectural principles can support individual, collaborative, learner-facing, and teacher-facing interventions within one coherent system.

When compared with representative related works, the proposed architecture shows several advantages. Unlike studies centered mainly on adaptive learning characteristics, the present work provides an explicit architectural model. Unlike analytics-focused studies, it includes agent-based orchestration and direct intervention delivery. Unlike LMS-integrated artificial intelligence assistants, it extends beyond a single assistant paradigm and supports a broader multi-agent ecosystem. Unlike function-specific recommendation or feedback systems, it is designed for heterogeneous LMS environments and incorporates standards-based interoperability as a central principle. Therefore, the novelty of this study lies not in the isolated use of one technology, but in the integration of interoperability standards, learning analytics, distributed intelligent agents, and generative artificial intelligence within a modular closed-loop smart learning architecture.

Despite these contributions, the study also has important limitations. First, the validation remains theoretical and scenario-based. Although this approach is appropriate for assessing functional completeness and architectural coherence at this stage, it does not provide empirical evidence regarding performance, latency, scalability, usability, or educational effectiveness. Second, the operational deployment of a distributed architecture combining analytics, intelligent agents, and generative artificial intelligence may introduce computational and governance challenges in real institutional settings. Third, although the framework includes conceptual safeguards for responsible artificial intelligence use, these mechanisms still need to be implemented and evaluated in authentic contexts. These limitations confirm that the present work should be understood as an architectural contribution and a foundation for future experimentation rather than as a fully validated deployment model.

Overall, the discussion confirms that the proposed framework provides a coherent, extensible, and interoperable contribution to smart learning research. Its main added value lies in the ability to connect standardized trace collection, learner modeling, intelligent agent orchestration, and pedagogically grounded interventions within a unified architecture that remains portable across heterogeneous LMS platforms. In this way, the study contributes not only to the conceptual design of next-generation smart learning environments but also to the definition of a reproducible architectural basis for future prototype development and empirical implementation.

VI. Conclusion

This study proposed an interoperable smart learning architecture designed to support adaptive and personalized learning across heterogeneous Learning Management System (LMS) environments. The work addressed the research gap identified in the literature by defining a modular framework that decouples intelligent services from the LMS while preserving standards-based interoperability. In this sense, the first objective of the study was achieved through the design of an architecture that separates the LMS hosting role from the intelligence layer and enables adaptive services to operate through interoperable protocols rather than through modifications of the LMS core.

The second objective of the study was to integrate learning analytics, intelligent agents, and generative artificial intelligence into a unified closed-loop adaptation process. The proposed framework achieved this by connecting trace collection, feature computation, learner modeling, agent-based orchestration, and intervention delivery within one coherent architecture. This integration represents one of the main contributions of the work because many existing approaches address these dimensions separately rather than as part of a coordinated educational ecosystem.

The third objective was to show how standardized learning traces can support both individual and collaborative pedagogical interventions. Through the four scenario-based validations, the study demonstrated that the same architectural logic can support early dropout-risk detection, targeted remediation after quiz failure, teacher-facing dashboards with grade return, and collaborative feedback for group learning situations. These scenarios highlight the functional completeness of the architecture and show its ability to support learner-facing, teacher-facing, individual, and collaborative services within the same framework.

The novelty of this work lies not in the isolated use of a single technology, but in the explicit integration of interoperability standards, learning analytics, distributed intelligent agents, and generative artificial intelligence within a modular and closed-loop smart learning architecture. Compared with narrower related works, the proposed framework places stronger emphasis on LMS independence, agent orchestration, pedagogically grounded analytics, and institutional portability across heterogeneous platforms. This gives the model a broader architectural scope and a stronger deployment perspective than many existing intelligent learning solutions.

From a scientific perspective, the study contributes a reproducible architectural foundation for smart learning research by bringing together research streams that are often treated separately. From a practical perspective, the framework offers institutions a structured basis for deploying adaptive services in existing LMS ecosystems while preserving governance, portability, and compatibility with standards-based infrastructures. In this way, the proposed model may support the future development of more scalable, flexible, and institutionally reliable smart learning environments.

At the same time, the study has important limitations. The validation remains theoretical and scenario-based, which means that the work does not yet provide empirical evidence regarding usability, performance, scalability, or educational effectiveness in authentic classroom settings. In addition, the implementation of intelligent agents and generative artificial intelligence in real institutional environments may raise technical, ethical, and governance-related challenges that require further investigation. These limitations indicate that the present study should be understood primarily as an architectural and conceptual contribution rather than a fully validated deployment model.

Future work should therefore focus on empirical implementation and evaluation in authentic educational contexts. This includes prototype development, experimentation in real LMS platforms, measurement of system performance and user experience, and assessment of the educational effectiveness of agent-based and generative artificial intelligence-supported interventions. Further research may also explore explainable and trust-aware agent behavior, more advanced multimodal learning analytics, and stronger governance mechanisms for privacy, ethics, and responsible artificial intelligence in smart learning systems.

REFERENCES

[1] Munna, M. S. H., Hossain, M. R., & Saylo, K. R. (2024). Digital education revolution: Evaluating LMS-based learning and traditional approaches. *Journal of Innovative Technology Convergence*, *6*(2). https://doi.org/10.69478/JITC2024v6n002a03

[2] Dehbi, A., Bakhouyi, A., Dehbi, R., & Talea, M. (2022). MOOCs in smart education: Comparative study by applying AHP and COPRAS method. *International Journal of Emerging Technologies in Learning (iJET)*, *17*(8), 61-74. https://doi.org/10.3991/ijet.v17i08.27871

[3] Abu-Rasheed, H., Weber, C., & Fathi, M. (2023, July). Context based learning: a survey of contextual indicators for personalized and adaptive learning recommendations–a pedagogical and technical perspective. In *Frontiers in Education* (Vol. 8, p. 1210968). Frontiers Media SA. https://doi.org/10.3389/feduc.2023.1210968

[4] Du Plooy, E., Casteleijn, D., & Franzsen, D. (2024). Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement. *Heliyon*, *10*(21). https://doi.org/10.1016/j.heliyon.2024.e39630

[5] Alalawi, K., Athauda, R., & Chiong, R. (2025). An extended learning analytics framework integrating machine learning and pedagogical approaches for student performance prediction and intervention. *International Journal of Artificial Intelligence in Education*, *35*(3), 1239-1287. https://doi.org/10.1007/s40593-024-00429-7

[6] Miller, B., Rutherford, T., Pack, A., & Johnson, A. (2021, November). Bridging the SCORM and xAPI Gap: The Role of cmi5. In *Interservice/Industry Training, Simulation, and Education Conference, November-December* (pp. 1-11).

[7] Rodríguez-Ortiz, M. Á., Santana-Mancilla, P. C., & Anido-Rifón, L. E. (2025). Machine learning and generative AI in learning analytics for higher education: A systematic review of models, trends, and challenges. *Applied Sciences*, *15*(15), 8679. https://doi.org/10.3390/app15158679

[8] Ngulube, P., & Ncube, M. M. (2025). Leveraging Learning Analytics to Improve the User Experience of Learning Management Systems in Higher Education Institutions. *Information*, *16*(5), 419. https://doi.org/10.3390/info16050419

[9] Samuelsen, J., Chen, W., & Wasson, B. (2021). Enriching context descriptions for enhanced LA scalability: a case study. *Research and Practice in Technology Enhanced Learning*, *16*(1), 6. https://doi.org/10.1186/s41039-021-00150-2

[10] Hariyanto, Kristianingsih, F. X. D., & Maharani, R. (2025). Artificial intelligence in adaptive education: a systematic review of techniques for personalized learning. *Discover Education*, *4*(1), 458. https://doi.org/10.1007/s44217-025-00908-6

[11] Sajja, R., Sermet, Y., Cwiertny, D., & Demir, I. (2025). Integrating AI and learning analytics for data-driven pedagogical decisions and personalized interventions in education. *Technology, knowledge and learning*, 1-31. https://doi.org/10.1007/s10758-025-09897-9

[12] Qian, Y. (2025). Pedagogical Applications of Generative AI in Higher Education: A Systematic Review of the Field. *TechTrends*, 1-16. https://doi.org/10.1007/s11528-025-01100-1

[13] Alier, M., Pereira, J., Garcia-Penalvo, F. J., Casan, M. J., & Cabre, J. (2025). LAMB: An open-source software framework to create artificial intelligence assistants deployed and integrated into learning management systems. *Computer Standards & Interfaces*, *92*, 103940. https://doi.org/10.1016/j.csi.2024.103940

[14] Tu, Y., Chen, J., & Huang, C. (2025). Empowering personalized learning with generative artificial intelligence: Mechanisms, challenges and

pathways. *Frontiers of Digital Education*, *2*(2), 1-18. https://doi.org/10.1007/s44366-025-0056-9

[15] Elsayed, H. (2024). The impact of hallucinated information in large language models on student learning outcomes: A critical examination of misinformation risks in AI-assisted education. *Northern Reviews on Algorithmic Research, Theoretical Computation, and Complexity*, *9*(8), 11-23. https://doi.org/10.3390/app15158679

[16] El Akhdar, A., Baidada, C., & Kartit, A. (2024, April). Adaptability of Microservices Architecture in IoT Systems: A Comprehensive Review. In *Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security* (pp. 1-9). https://doi.org/10.1145/3659677.3659734

[17] Dehbi, A., Bakhouyi, A., Dehbi, R., & Talea, M. (2024, December). Towards a Smart Learning Environment: A Model for Smart Education using IoT and AI. In *2024 6th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)* (pp. 1-5). IEEE. https://doi.org/10.1109/ISAECT64333.2024.10799743

[18] Dragoni, N., Giallorenzo, S., Lafuente, A. L., Mazzara, M., Montesi, F., Mustafin, R., & Safina, L. (2017). Microservices: yesterday, today, and tomorrow. *Present and ulterior software engineering*, 195-216. https://doi.org/10.1007/978-3-319-67425-4_12

[19] Moreschini, S., Pour, S., Lanese, I., Balouek, D., Bogner, J., Li, X., ... & Taibi, D. (2025). AI Techniques in the Microservices Life-Cycle: a Systematic Mapping Study. *Computing*, *107*(4), 100. https://doi.org/10.1007/s00607-025-01432-z

[20] Tan, L. Y., Hu, S., Yeo, D. J., & Cheong, K. H. (2025). Artificial Intelligence-Enabled Adaptive Learning Platforms: A Review. *Computers and Education: Artificial Intelligence*, 100429. https://doi.org/10.1016/j.caeai.2025.100429

[21] Strielkowski, W., Grebennikova, V., Lisovskiy, A., Rakhimova, G., & Vasileva, T. (2025). AI-driven adaptive learning for sustainable educational transformation. *Sustainable Development*, *33*(2), 1921-1947. https://doi.org/10.1002/sd.3221

[22] Molinari, A., & Sandri, S. (2024, November). Evolution of lms design and implementation in the age of ai and large language models. In *Proceedings of the Second International Workshop on Artificial INtelligent Systems in Education co-located with 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024), Bolzano, Italy*.

[23] Bakhouyi, A., Dehbi, A., Amhaimar, L., Broumi, S., Talea, M., & Khalidi, A. (2025). Improving SMART learning: Course completion via AI-driven hybrid system integration in big data. Telematics and Informatics Reports, 18, 100199. https://doi.org/10.1016/j.teler.2025.100199

[24] Khaddar, A. M., Dehdi, A., Alali, A., Bakhouyi, A., & Talea, M. (2025). Emerging Technologies in Learning: A Bibliometric Analysis of Technology Integration and Applications. International Journal of Interactive Mobile Technologies, 19(5). https://doi.org/10.3991/ijim.v19i05.51267

[25] Gilbert, N., & Troitzsch, K. (2005). Simulation for the social scientist. McGraw-Hill Education (UK).

[26] Railsback, S. F., & Grimm, V. (2019). *Agent-based and individual-based modeling: a practical introduction*. Princeton university press.

[27] Dehbi, A., Bakhouyi, A., Dehbi, R., & Talea, M. (2024). Smart evaluation: a new approach improving the assessment management process through cloud and IoT technologies. *International journal of information and education technology*, *14*(1), 107-118. https://doi.org/10.18178/ijiet.2024.14.1.2030

[28] Yan, L., Echeverria, V., Jin, Y., Fernandez-Nieto, G., Zhao, L., Li, X., ... & Martinez-Maldonado, R. (2024). Evidence-based multimodal learning analytics for feedback and reflection in collaborative learning. *British Journal of Educational Technology*, *55*(5), 1900-1925. https://doi.org/10.1111/bjet.13498

[29] Bakhouyi, A., Dehbi, A., Amhaimar, L., Tazouti, Y., Nadir, Y., & Khalidi, A. (2025). Enhancing learning outcomes in smart education: A supervised machine learning predictive analytics model for course completion. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 14(6), 4711–4721. https://doi.org/10.11591/ijai.v14.i6.pp4711-4721

[30] Nongkhai, L. N., Wang, J., Wynn, A., & Mendori, T. (2025). Evaluating adaptive and generative AI-based feedback and recommendations in a knowledge-graph-integrated programming learning system. *Computers and Education: Artificial Intelligence*, 100526. https://doi.org/10.1016/j.caeai.2025.100526

[31] Mazzullo, E., Bulut, O., Wongvorachan, T., & Tan, B. (2023). Learning analytics in the era of large language models. *Analytics*, *2*(4), 877-898. https://doi.org/10.3390/analytics2040046

[32] Dehbi, A., Bakhouyi, A., Dehbi, R., & Talea, M. Towards Technology-Enhanced Learning: A Novel Machine Learning Approach in Education 4.0. *Available at SSRN 4974380*. https://doi.org/10.1016/j.teler.2025.100194

Mapping Tourist Sentiments Through Lexicon-Based Analysis of Social Media Reviews: The Case of Salak Sibetan Agritourism

Paula Dewanti[1], Putu Adi Guna Permana[2], I Gusti Ayu Widari Upadani[3],
Ellyn Ly Maramento[4], Alfred John G. Borreros[5]
Department of Informatics and Computer, Institut Teknologi dan Bisnis STIKOM Bali, Denpasar, Indonesia [1, 2, 3]
College of Tourism and Hospitality Management, Guimaras State University, Guimaras, Phillipines [4]
College of Science and Technology, Guimaras State University, Guimaras, Phillipines [5]

***Abstract*—Salak Sibetan, Bali's emblematic snake fruit cultivated in Sibetan Village, Karangasem, has gained increasing digital visibility through user-generated content across social media platforms. This study applies a bilingual lexicon-based sentiment analysis framework integrating Indonesian and English sentiment lexicons, explicit negation handling, domain-specific agritourism vocabulary, and a mean-based sentiment scoring function to classify 500 online reviews collected from Facebook, Instagram, Shopee, TikTok, and Twitter/X. The methodology adapts Indonesian and English lexicons with negation handling, domain-specific refinement for agritourism terminology, and mean-based scoring to improve neutrality discrimination. Results indicate an overall sentiment distribution of 40% positive, 40% neutral, and 20% negative. Positive reviews emphasize taste quality (*manis legit*, *renyah*, fresh) and cultural authenticity, while negative feedback highlights packaging issues, inconsistent quality, and pricing concerns. Image-centric platforms (Instagram and TikTok) exhibit higher proportions of positive sentiment emphasizing taste quality and authenticity, whereas transaction-oriented platforms (Shopee and Twitter/X) show more neutral and negative expressions related to logistics, packaging, and pricing. Beyond sentiment measurement, the study demonstrates how lexicon-based methods can capture platform-specific evaluative behavior within heritage agritourism contexts, offering methodological insights for multilingual sentiment analysis in low-resource domains and strategic implications for sustainable destination communication aligned with Sibetan's FAO GIAHS recognition.**

***Keywords*—Sentiment analysis; lexicon-based methods; Salak Sibetan; agritourism; GIAHS; social media analytics**

I. Introduction

Tourism is one of the most rapidly expanding sectors in Indonesia, contributing approximately 4.3% to the national GDP in 2022 [1]. Within this national context, Bali remains Indonesia's premier tourism destination, where agritourism has emerged as an important mechanism for rural economic development and cultural preservation. Among Bali's distinctive agricultural heritage areas, Sibetan Village in Karangasem Regency is widely recognized for the cultivation of Salak Sibetan (Salacca zalacca var. amboinensis), a snake fruit variety valued for its unique sweetness, texture, and cultural significance [2].

The growing influence of digital platforms such as Instagram, Facebook, TikTok, and e-commerce marketplaces like Shopee has fundamentally transformed how agritourism destinations and agricultural products are evaluated and promoted. User-Generated Content (UGC), including online reviews and social media posts, functions as electronic word-of-mouth that shapes visitor expectations, purchasing decisions, and destination image. In 2024, the Food and Agriculture Organization (FAO) designated the Salak Sibetan production system as a Globally Important Agricultural Heritage System (GIAHS), reinforcing its ecological and cultural value and strengthening its potential positioning as a heritage-based agritourism destination [3].

Sentiment analysis has been widely adopted in tourism and hospitality research to interpret large volumes of UGC and extract insights into visitor satisfaction, destination image, and service quality. Prior studies have applied sentiment analysis to hotels, urban destinations, and mass tourism attractions using both machine learning and lexicon-based approaches. However, important limitations remain in existing literature. First, agritourism, particularly heritage-based agricultural systems in Indonesia, has received far less analytical attention than urban or conventional tourism destinations. Second, many existing studies focus on monolingual datasets, despite the prevalence of bilingual and code-mixed Indonesian–English content on social media platforms. Third, cross-platform sentiment analyses are often presented descriptively, without sufficient consideration of how platform characteristics shape sentiment expression. This study addresses these gaps by applying a bilingual, domain-adapted lexicon-based sentiment framework across multiple social media and e-commerce platforms to examine how perceptions of Salak Sibetan agritourism differ by platform context.

This study seeks to explore how Salak Sibetan agritourism is perceived across multiple social media platforms by examining the dominant positive, neutral, and negative sentiments expressed in visitor and consumer narratives. It aims to: 1) apply a lexicon-based sentiment analysis approach to bilingual (Indonesian-English) reviews, 2) identify key sentiment patterns and recurring themes across platforms, 3) analyze platform-specific differences in sentiment expression, and 4) translate these insights into practical

recommendations for improving product quality, service delivery, and digital communication strategies.

The significance of this study is twofold. From an academic perspective, it extends sentiment analysis research to the underexplored domain of Indonesian heritage agritourism, demonstrating the suitability of lexicon-based methods for multilingual and low-resource contexts while highlighting the analytical significance of neutral sentiment. From a practical standpoint, the findings provide guidance for farmers, local communities, tourism authorities, and e-commerce sellers seeking to enhance visitor experiences, address quality concerns, and strengthen digital engagement for Salak Sibetan as a premium agritourism product.

II. Literature Review

A. Social Media and Agritourism

Social media has become a central arena for travelers and consumers to exchange real-time photos, narratives, and evaluations of agritourism products and experiences. This user-generated content drives electronic word-of-mouth, shaping information search, destination visibility, and purchasing intentions [4]. For agritourism settings such as Salak Sibetan, e-WOM not only communicates product attributes (taste, freshness, value) but also circulates stories about local culture and on-site experiences, thereby influencing perceptions of authenticity and community engagement. Positive reviews can strengthen brand salience and trust, whereas negative feedback may signal service gaps and operational risks for destination managers [5]. Fig. 1 shows the TikTok hashtag for Salak Sibetan.

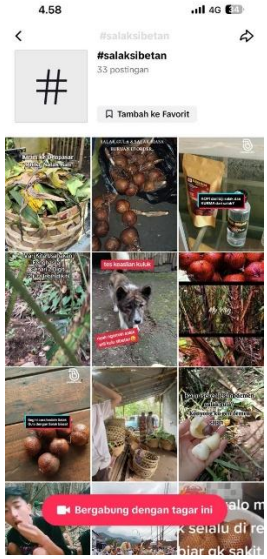


Fig. 1. TikTok Hashtag Salak Sibetan.

B. Sentiment Analysis for Tourism Analytics

Sentiment analysis provides a scalable means to interpret large volumes of UGC by classifying evaluative language into positive, neutral, or negative polarity. In tourism analytics, this approach has been widely used to extract strategic insights from large-scale UGC, enabling destination managers to understand collective visitor perceptions, as demonstrated in urban destination studies such as Barcelona [6]. Previous tourism applications have focused on general destinations or hospitality services using various computational approaches [7]. Lexicon-based methods offer interpretability and low-resource advantages for multilingual contexts, making them suitable for Indonesian-English mixed content common in social media [8]. Studies have demonstrated that lexicon-based approaches with negation handling and domain-specific refinement can effectively capture sentiment in Indonesian tourism reviews [9].

Research on sentiment analysis for agricultural products and agritourism remains limited. Pratiwi et al. [10] conducted lexicon-based sentiment analysis of social media reviews for floating markets in South Kalimantan, demonstrating the applicability of such methods to Indonesian heritage tourism contexts. Their findings highlighted the importance of neutral reviews in providing factual information and the value of cross-platform analysis for understanding different audience segments. In parallel, recent tourism studies have employed machine learning-based sentiment analysis of TripAdvisor reviews to uncover experiential patterns and visitor preferences, underscoring the growing role of online review platforms in tourism intelligence [11]. Destination-focused sentiment analysis has also been conducted using TripAdvisor data to model visitor experiences and perceptions, as illustrated by studies of heritage destinations such as Dubrovnik [12].

C. Salak Sibetan as Competitive Agritourism

Salak Sibetan, cultivated in Sibetan Village, Karangasem, is among Indonesia's most distinctive agritourism commodities. Its competitive advantage arises from a blend of agro-biodiversity, cultural heritage, and growing digital visibility [2]. The village hosts numerous salak varieties ($\geqslant 14$), including premium cultivars such as gula pasir, celebrated for unique sweetness and texture. This varietal richness and a five-strata agroforestry system embedded in local customary rules and Tri Hita Karana philosophy underpin both ecological resilience and visitor appeal [3]. Recent advances in smart agritourism demonstrate how digital marketing and information systems help rural destinations amplify reach, structure content, and convert interest into visits. Empirical work shows that UGC and e-WOM significantly influence destination image and tourist visit intention, particularly on platforms like Instagram, Google Reviews, and TikTok [11]. In Sibetan, government and community profiles highlight orchard experiences, salak tasting, salak-derived products (coffee, syrup, wine), and eco-cultural treks as core narratives increasingly promoted via social channels.

D. Research Gap and Contribution

This study addresses several gaps in existing literature. First, it provides the first comprehensive sentiment analysis specifically focused on Salak Sibetan agritourism across multiple social media platforms. Second, it employs a bilingual lexicon-based approach with domain-specific refinement for agritourism terminology, including taste descriptions, Salak varieties, and quality indicators. Third, it examines cross-platform differences in sentiment expression, comparing image-heavy platforms (Instagram, TikTok) with e-commerce (Shopee) and microblogging (Twitter/X) platforms. Fourth, it highlights the role of neutral reviews in providing credibility-

enhancing information about product origin, logistics, and practical visitor guidance. While recent studies have introduced advanced approaches such as aspect-based sentiment analysis and zero-shot learning for hospitality service enhancement [13], these methods often require substantial computational resources and annotated data, making them less suitable for low-resource agritourism contexts.

III. Research Methods

A. Research Design

This study employs a quantitative content analysis to examine user-generated content related to Salak Sibetan agritourism. The methodological workflow consists of five stages: data collection, preprocessing, sentiment classification, visualization, and validation. Each stage was designed to ensure analytical rigor, reproducibility, and alignment with established standards in computational tourism analytics [14].

B. Data Collection

The dataset comprised 500 online reviews collected from five widely used social media platforms: Facebook (106 reviews; 21.2%), Instagram (103; 20.6%), Shopee (100; 20.0%), TikTok (103; 20.6%), and Twitter/X (88; 17.6%). All reviews were written in Indonesian, reflecting feedback from local and domestic visitors and consumers. The average review length was 28.58 words. Fig. 2 shows an example review from a TikTok user.



Fig. 2. Example of TikTok User.

Reviews were included if they contained at least one coherent sentence describing Salak Sibetan products, experiences, or purchase interactions. Data were collected using keyword searches for "salak sibetan," "salak sibetan karangasem," and "salak bali" across platform search functions during October-December 2024.

C. Preprocessing

The preprocessing stage consisted of several steps designed to prepare the review text for analysis. The process began with text normalization, including converting all text to lowercase, applying Unicode standardization, and removing unwanted noise such as stray symbols or characters. After that, stop-word removal and selective slang expansion were performed using bilingual stop-word lists to ensure clearer semantic interpretation [15]. Stemming or lemmatization was applied, when necessary, particularly for generating keyword-based summaries. As a final step, manual checks were conducted to confirm that the preprocessing workflow preserved the original meaning of the content.

D. Sentiment Classification

For sentiment analysis, a lexicon-based approach was applied to categorize reviews into positive, neutral, or negative sentiments. The study adapted two existing lexicons: InSet for Indonesian and NRC for English, with domain-specific refinement adding tourism and agritourism terminology such as taste descriptions (manis legit, renyah, sepet), salak varieties (gula pasir), quality indicators (fresh, lembek, memar), and logistics terms (packing, pengiriman). Each token was assigned a polarity score (+1, -1, or 0), with negation handling implemented for words preceded by terms such as tidak, bukan, or not. The review's overall sentiment was determined by the aggregate mean score of its tokens, with thresholds: >0.1 positive, < -0.1 negative, and between -0.1 and 0.1 neutral. Mean-based scoring was selected over summation to improve neutrality discrimination and reduce bias from review length [16].

$$S(r)=\frac{1}{n}\sum_{i=1}^{n} w_i \quad (1)$$

Variable Description:

- S(r): The total sentiment score of review r.
- n: The total number of tokens (words) in the review that match entries in the sentiment lexicon.
- w_i: The polarity value of the i-th word, where +1 indicates positive sentiment, −1 indicates negative sentiment, and 0 indicates neutral sentiment.

E. Validation

To ensure the reliability of the sentiment classification, a manual validation procedure was conducted. A subset of 50 reviews (10% of dataset) was independently annotated by two human coders with expertise in tourism communication and text analysis. The degree of agreement between the coders was assessed using Cohen's Kappa, which yielded a coefficient of 0.82, indicating strong inter-rater reliability. This validation step confirmed that the lexicon-based model performed consistently with human judgment, thereby supporting the credibility of the analytical results.

F. Data Analysis

The classified sentiment outcomes were subsequently examined through descriptive and visual analytics. Bar charts and pie charts were produced to illustrate the distribution of sentiment categories. Platform-based differences were analyzed using comparative tables. Keyword frequency analysis identified recurring terms in each sentiment category, highlighting dominant themes and concerns.

IV. Results and Discussion

A. Dataset Characteristics

The dataset comprised 500 online reviews distributed across five platforms, as shown in Table I. Facebook contributed the largest share (21.2%), followed by Instagram (20.6%), TikTok (20.6%), Shopee (20.0%), and Twitter/X (17.6%). Average review length was consistent across platforms, ranging from

27.10 words on Instagram to 30.30 words on Twitter/X, indicating similar levels of detail in user-generated content regardless of platform. Fig. 3 shows the sentiment distribution across five social media platforms.

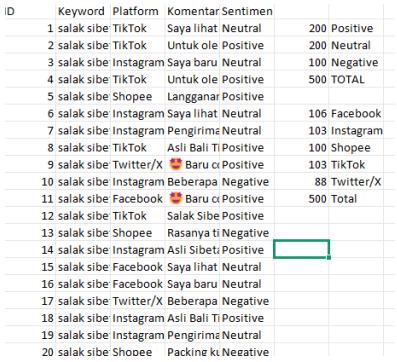


Fig. 3. Sentiment distribution across five social media platforms.

TABLE I. DATASET CHARACTERISTICS BY PLATFORM

Platform	Reviews	Percentage	Avg. Length (words)
Facebook	106	21.2%	29.40
Instagram	103	20.6%	27.10
Shopee	100	20.0%	28.20
TikTok	103	20.6%	27.90
Twitter/X	88	17.6%	30.30
Total	**500**	**100%**	**28.58**

B. Overall Sentiment Distribution

Sentiment classification divided the dataset into three categories: 200 reviews (40%) positive, 200 reviews (40%) neutral, and 100 reviews (20%) negative, as presented in Table II. The dominance of positive sentiment indicates that Salak Sibetan is generally well-received by consumers across multiple platforms. At the same time, the equally high proportion of neutral reviews suggests that a substantial number of users provide objective or descriptive comments, focusing on factual product information rather than emotional expressions. The presence of 20% negative reviews highlights areas that may require managerial attention, as unfavorable consumer experiences shared online can significantly influence future purchasing behavior and digital engagement.

TABLE II. OVERALL SENTIMENT DISTRIBUTION

Sentiment	Frequency	Percentage
Positive	200	40%
Neutral	200	40%
Negative	100	20%
Total	500	100%

C. Sentiment by Platform

The sentiment analysis was disaggregated to compare results across the five platforms, revealing important differences in how consumers express their perceptions, shaped by the nature of each platform. Table III presents the sentiment distribution by platform.

TABLE III. SENTIMENT DISTRIBUTION BY PLATFORM

Platform	Positive	Neutral	Negative	Total
Facebook	42 (39.6%)	44 (41.5%)	20 (18.9%)	106
Instagram	47 (45.6%)	36 (35.0%)	20 (19.4%)	103
Shopee	38 (38.0%)	42 (42.0%)	20 (20.0%)	100
TikTok	44 (42.7%)	41 (39.8%)	18 (17.5%)	103
Twitter/X	29 (33.0%)	37 (42.0%)	22 (25.0%)	88
Total	**200**	**200**	**100**	**500**

Instagram shows the highest proportion of positive sentiment (45.6%), likely because visual platforms encourage sharing of aesthetically pleasing product photos and positive experiences. TikTok follows closely with 42.7% positive, reflecting the platform's trend-driven nature where users share recommendations and product highlights. Shopee exhibits balanced sentiment with 38% positive, 42% neutral, and 20% negative, characteristic of e-commerce platforms where consumers provide both praise and critical feedback about product quality and delivery.

Twitter/X shows the lowest positive sentiment (33.0%) and highest negative sentiment (25.0%), possibly due to the platform's use for expressing complaints and seeking resolution. The high neutral proportion on Twitter/X (42.0%) and Shopee (42.0%) indicates that users on these platforms frequently share factual information about product availability, pricing, and logistics.

D. Keyword and Thematic Analysis

Keyword analysis identified recurring terms in each sentiment category, revealing dominant themes in consumer discourse about Salak Sibetan. Positive reviews emphasized words such as manis legit (sweet), renyah (crispy), fresh, recommended, top, juara (champion), kriuk (crunchy), and padat (dense). Many highlighted the distinctive taste quality of Salak Sibetan, particularly the gula pasir variety, and its suitability as a souvenir (oleh-oleh) from East Bali. Examples include: "Asli Bali Timur, salak Sibetan ini top—tidak sepet, renyah" (Authentic East Bali, this Salak Sibetan is top—not astringent, crispy) and "Untuk oleh-oleh Bali Timur, salak Sibetan Karangasem ini juara sih" (For East Bali souvenirs, this Salak Sibetan Karangasem is champion). Neutral reviews commonly contained words such as asal (origin), Desa Sibetan, Karangasem, Bali Timur, pengiriman (delivery), jadwal (schedule), pertama kali (first time), and cari (search). These were descriptive and provided practical information about product origin, delivery experiences, and initial encounters. Examples include: "Salak Sibetan itu berasal dari Desa Sibetan" (Salak Sibetan comes from Sibetan Village) and "Saya baru pertama kali mencoba salak Sibetan Karangasem" (I tried Salak Sibetan Karangasem for the first time).

Negative reviews frequently included words like packing (packaging), *lembek* (soggy), *memar* (bruised), *rusak* (damaged), *kurang manis* (less sweet), *sepet* (astringent), and *harga* (price). These highlight operational problems such as inadequate packaging during shipping, inconsistent fruit quality, and pricing concerns. Examples include: "*Packing kurang tebal, jadi ada buah yang rusak*" (Packaging was not thick enough, so

some fruits were damaged) and "*Harga salak sibetan karangasem menurut saya agak mahal*" (In my opinion, the price of Salak Sibetan Karangasem is somewhat expensive).

E. Interpretation of Neutral Reviews

Neutral reviews, which accounted for 40% of the dataset, primarily contained factual and descriptive information rather than emotional expressions. Visitors and consumers often mentioned elements such as product origin, first-time experiences, delivery schedules, or general observations without attaching strong judgments. For instance, one reviewer wrote, "*Saya lihat banyak yang cari salak sibetan karangasem akhir-akhir ini*" (I see many people searching for Salak Sibetan Karangasem lately), while another stated, "*Pengiriman salak Sibetan Karangasem sampai sesuai jadwal*" (Delivery of Salak Sibetan Karangasem arrived on schedule). These examples show how neutral reviews serve as practical guides that help potential buyers understand product availability and logistics.

The relatively high proportion of neutral reviews contributes to the credibility and authenticity of online platforms. While positive and negative reviews reflect emotions, neutral reviews provide balance by offering straightforward accounts of products and experiences. Their stability across platforms indicates that they function as a baseline narrative, consistently delivering reliable context regardless of variations in sentiment trends. From a managerial perspective, these reviews highlight recurring topics about origin and logistics, signaling that clearer communication in promotional materials could enhance consumer understanding and trust.

F. Negative Sentiment and Service Gaps

Negative reviews, although representing only 20% of the dataset, carry significant weight in shaping overall perception of Salak Sibetan products. These reviews frequently pointed out recurring problems such as inadequate packaging leading to damaged fruit, inconsistent taste quality, and perceptions of high prices. For example, one consumer noted, "Packing kurang rapi, jadi ada buah yang memar" (Packaging was not neat, so some fruits were bruised), while another commented, "*Rasanya kurang istimewa, tidak sesuai yang saya bayangkan*" (The taste was less special, not as I imagined). Such feedback highlights those operational issues, particularly in e-commerce fulfillment, are the main drivers of dissatisfaction.

From a managerial perspective, these negative reviews reveal critical service gaps that demand immediate attention. Packaging quality must be prioritized to preserve product condition during shipping, while quality control measures should ensure consistent taste and freshness. Addressing concerns about pricing transparency is also essential, as perceptions of high prices without corresponding quality can damage trust and deter repeat purchases. If left unresolved, these issues risk eroding the long-term competitiveness of Salak Sibetan in both e-commerce and agritourism markets.

G. Social Media as Double-Edged Sword

Social media operates as a dual-impact mechanism in shaping the digital reputation of Salak Sibetan agritourism. On the positive side, large volumes of favorable and neutral user-generated content function as organic promotion. Reviews and images emphasizing taste quality, authenticity, and suitability as souvenirs amplify product visibility across platforms such as Instagram, TikTok, Shopee, and Facebook. These narratives significantly influence consumer purchasing decisions and act as low-cost digital marketing channels [17].

Conversely, negative reviews spread rapidly and can harm reputation if ignored. Complaints about packaging damage, inconsistent quality, and pricing easily circulate online, discouraging potential buyers. Continuous monitoring through automated sentiment analysis enables managers to balance positive promotion with early warnings, ensuring adaptive strategies and sustainable competitiveness in digital agritourism.

H. Managerial Implications

The findings of this study provide actionable insights for integrating sentiment analysis into agritourism and e-commerce management. First, the predominance of positive reviews about taste quality and authenticity suggests opportunities for leveraging data-driven promotional strategies. Destination managers and online sellers can amplify user-generated narratives across social media platforms, particularly emphasizing the GIAHS recognition and cultural heritage aspects that differentiate Salak Sibetan from ordinary snake fruit. Second, the recurrent negative sentiments regarding packaging, inconsistent quality, and pricing indicate specific service quality gaps that can be addressed using data-informed decision-making. Sellers should invest in better packaging materials, implement quality control procedures, and communicate transparently about seasonal price variations. Third, the high proportion of neutral reviews highlights the importance of providing clear factual information about product origin, delivery schedules, and harvesting seasons in promotional materials. Fourth, the adoption of sentiment monitoring tools would allow sellers and destination managers to detect emerging issues proactively and respond to customer concerns. This approach demonstrates how computational sentiment analysis can function as a decision-support mechanism, positioning heritage products like Salak Sibetan as benchmarks for quality in Indonesian agritourism.

I. Theoretical Implications

This study contributes theoretically by extending sentiment analysis to agritourism and heritage food products in Indonesia. Using lexicon-based methods, it demonstrates how computational models capture cultural, quality, and logistics dynamics in consumer discourse. Consistent with previous research on online destination image, positive and negative narratives shared through social media significantly influence tourist engagement and perception, reinforcing the strategic importance of monitoring UGC sentiment [18]. The significant share of neutral reviews reveals that e-WOM for agricultural products is both emotional and informational, with neutral content serving as a stabilizing factor for credibility. Additionally, differences identified across platforms underline the necessity of adopting cross-platform approaches in sentiment analysis. The contrast between visual platforms (Instagram, TikTok) showing higher positive sentiment and e-commerce platforms (Shopee) showing balanced expressions indicates that platform characteristics shape sentiment expression. These insights refine sentiment analysis theory by

emphasizing contextual and platform-specific factors in agritourism-focused applications

V. Conclusions and Future Work

Rather than merely quantifying sentiment polarity, this study demonstrates how lexicon-based sentiment analysis can function as an interpretable analytical tool for heritage agritourism contexts characterized by multilingual, platform-diverse user-generated content. The findings illustrate that sentiment expression is not uniform across platforms but shaped by platform affordances and user intent, underscoring the importance of cross-platform designs in sentiment analysis research. Moreover, the substantial presence of neutral sentiment reveals the informational dimension of electronic word-of-mouth, which plays a critical yet often overlooked role in credibility building and destination understanding.

This study has several limitations that should be acknowledged. The dataset was restricted to Indonesian-language reviews, which may not represent perceptions of international visitors. The lexicon-based method, although transparent and suitable for multilingual data, may not fully capture nuanced expressions such as sarcasm or implicit sentiment. The analysis also relied solely on textual content, excluding visual cues that influence consumer evaluations on image-heavy platforms. Furthermore, the cross-sectional design does not capture temporal sentiment shifts across harvest seasons. Future research may extend this study by incorporating larger and more diverse datasets from additional platforms, such as Google Reviews and TripAdvisor, and by integrating advanced techniques, including aspect-based sentiment analysis or zero-shot sentiment models, to capture more nuanced experiential dimensions beyond the capabilities of lexicon-based approaches. Employing advanced machine learning methods, including transformer-based deep learning models such as IndoBERT, could improve classification accuracy and capture nuanced emotional expressions beyond lexicon constraints. Longitudinal analyses tracking sentiments across harvest seasons would enable more robust trend detection, while cross-cultural comparative studies could examine how international visitors perceive Salak Sibetan compared to domestic consumers. Finally, incorporating multimodal analysis including images and videos from platforms like Instagram and TikTok would provide a richer understanding of how visual content shapes product perceptions in agritourism.

References

[1] Badan Pusat Statistik, "Statistik Wisatawan Mancanegara 2023," BPS, Jakarta, 2024.

[2] I Nengah Suaningsih, Ni Luh Erni Puspita Utami, dan I Wayan Supartha, "Agritourism development strategy in Sibetan Village, Karangasem Regency, Bali," Journal of Tourism and Hospitality, vol. 8, no. 2, pp. 89-97, 2023.

[3] Food and Agriculture Organization, "Globally Important Agricultural Heritage Systems (GIAHS): Salak Sibetan System, Indonesia," FAO, Rome, 2024.

[4] Stephen W. Litvin, Ronald E. Goldsmith, dan Bing Pan, "Electronic word-of-mouth in hospitality and tourism management," Tourism Management, vol. 29, no. 3, pp. 458-468, 2008.

[5] Zheng Xiang dan Ulrike Gretzel, "Role of social media in online travel information search," Tourism Management, vol. 31, no. 2, pp. 179-188, 2010.

[6] Marine Elshatshat dan Shao Ah Cheng, "Tourism analytics with massive user-generated content: A case study of Barcelona," Journal of Destination Marketing & Management, vol. 18, p. 100469, 2020.

[7] Bing Liu, "Sentiment Analysis and Opinion Mining," San Rafael: Morgan & Claypool Publishers, 2012.

[8] Kalliopi Chatzilaou, Angelos Liapis, dan Michalis Hanas, "Sentiment Analysis for Tourism Insights: A Machine Learning Approach to TripAdvisor Reviews," Informatics, vol. 7, no. 4, p. 90, 2024.

[9] Fernando Rosado dan Maria Fillieri, "E-WOM and accommodation: An analysis of the factors that influence travelers' adoption of information," Journal of Travel Research, vol. 53, no. 1, pp. 44-57, 2014.

[10] Pratiwi, E. L., Pratama, R. N., Ahyati, I. U., & Dewanti, P. (2025). Lexicon-Based Sentiment Analysis of Social Media Reviews for Floating Market Popularity in South Kalimantan. *International Journal of Advanced Computer Science & Applications*, *16*(11).

[11] Talukder, S. C. (2025). Exploring the world: A journey through culture and nature. *Journal of Educational Studies*, *3*(1), 1-10.

[12] Ivana Zrinjski, Franka Šmit-Meić, Hrvoje Mesić, dan Boris Bosančić, "A sentiment analysis model based on user experiences of Dubrovnik on the TripAdvisor platform," Applied Sciences, vol. 14, no. 18, p. 8304, 2024

[13] Nadia Imami, Indah Kristiyani, Faisal, Muhammad Muzakir, Ridwan, dan Sani Muhammad, "Exploring tourist experience through online reviews using aspect-based sentiment analysis with zero-shot learning for hospitality service enhancement," Information, vol. 15, no. 8, p. 499, 2024.

[14] Le, T. H., Arcodia, C., Novais, M. A., & Kralj, A. (2021). Proposing a systematic approach for integrating traditional research methods into machine learning in text analytics in tourism and hospitality. *Current Issues in Tourism*, *24*(12), 1640-1655.Beverley A. Sparks dan Victoria Browning, "The impact of online reviews on hotel booking intentions and perception of trust," Tourism Management, vol. 32, no. 6, pp. 1310-1323, 2011.

[15] Aliero, A. A., Bashir, S. A., Aliyu, H. O., Tafida, A. G., Bashar, U. K., & Nasiru, M. D. (2023). Systematic review on text normalization techniques and its approach to non-standard words.

[16] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, dan Manfred Stede, "Lexicon-based methods for sentiment analysis," Computational Linguistics, vol. 37, no. 2, pp. 267-307, 2011.

[17] Adebayo Dayo, Adekunle Taiwo, Sunday Eyitayo, Hammed Adeola, dan Shola Ojo, "Sentiment analysis for assessment of hotel services review using feature selection approach based on decision tree," International Journal of Advanced Computer Science and Applications, vol. 11, no. 3, pp. 240-245, 2020.

[18] Mariana Gouveia, "The online destination image as portrayed by the user-generated content on social media and its impact on tourists' engagement," Tourism & Management Studies, vol. 20, no. 2, pp. 45-58, 2024

Optimization of Access Point Location Using an Integer Programming Model

Joan Marie M. Panes, Marilou O. Espina, Jovelin M. Lapates
Bukidnon State University, Malaybalay City, Bukidnon, Philippines

***Abstract*—To address the increasing demand for reliable connectivity in educational environments, strategic access point (AP) placement is necessary to support coverage and bandwidth requirements while controlling installation cost. This study applies a binary integer programming model to the third floor of the College of Technologies at Bukidnon State University in order to determine a feasible minimum-cost AP configuration under site-specific conditions. The model incorporates coverage, bandwidth, range, and AP-count constraints, together with demand estimates based on expected simultaneous network use. Results show that the optimization framework can identify a feasible AP-to-area assignment that satisfies the imposed constraints within the study site. Because the nominal AP ranges are large relative to the dimensions of the floor, the resulting configuration is influenced more strongly by installation cost and bandwidth feasibility than by range limitation alone. The findings show the practical value of binary integer programming as a structured planning tool for access point placement in bounded indoor environments. While the resulting configuration is specific to the physical, technical, and cost conditions of the study area, the modeling approach may be applied to other institutional settings through appropriate parameter recalibration. A post-deployment wireless site survey is recommended to verify actual field performance and identify any remaining coverage gaps.**

***Keywords*—Access points (APs); optimization; wireless network; coverage; integer programming**

I. Introduction

Reliable Wi-Fi service in educational buildings depends on effective access point placement that can satisfy coverage and bandwidth needs without unnecessary installation cost. Bukidnon State University's College of Technologies Annex B Campus already has a documented network design proposal, which establishes the institutional context for wireless planning in the same setting [1], [2].

In [1], the authors discuss wireless propagation across diverse environments and explain how physical barriers affect signal attenuation. In [3] and [4], the authors also examine the effects of buildings, vegetation, and terrain on signal strength and coverage, while [5] analyzes diffraction, reflection, and scattering in obstacle-rich environments. In [6], [7], and [8], the authors discuss applications of linear programming and the simplex algorithm in optimization, while [9] and [10] address path loss modeling. In [11], the authors examine user density in wireless mesh networks, [12] compares propagation across the 2.4 GHz and 5 GHz bands, [13] and [14] analyze broader coverage models, and [15] and [16] discuss deployment cost considerations.

Designing a network requires careful consideration of several factors. In [17], the authors emphasize the importance of determining the optimum number of APs for reliable performance in wireless communication systems. Indoor studies and optimization work show that adequate WLAN coverage can be achieved with a minimal number of strategically placed APs ([18]; [19]). However, determining the number and optimal placement of APs in indoor or outdoor environments is challenging. Effective design must account for adequate coverage, minimum signal strength at all nodes, required bandwidth, and the presence of obstructions, reflections, and signal interference. Such a design demands proper modeling and formulation of the problem as an optimization task with multiple constraints.

Recent WLAN planning continues to treat AP deployment as a constrained optimization problem where placement and configuration determine coverage quality, interference risk, and total deployment cost [20].

Genetic Algorithms (GA) remain used in Wi-Fi AP deployment problems because they can search large, discrete location spaces using fitness-driven evolution [22]. However, GA can be computationally demanding because it repeatedly evaluates a population of candidate solutions across many generations [23].

Particle Swarm Optimization (PSO) is also applied to indoor AP placement, including binary PSO variants that select AP locations under deployment constraints [24]. PSO is often preferred for its simple implementation and relatively few parameters to tune ([25]; [26]). Despite this, PSO can suffer from local-optima trapping and loss of swarm diversity, which can reduce solution quality on complex search spaces [25].

Integer programming remains a strong method for AP placement because it supports binary site-selection decisions and explicit constraints for coverage, bandwidth, and cost [27]. Recent work also combines placement with configuration decisions such as channel assignment and transmit power within ILP-based formulations [29]. Although the solve time increases as problem size grows, integer programming can produce optimal solutions for moderate-sized deployment instances [27].

This study applies a binary integer programming model to the third floor of the College of Technologies at Bukidnon State University. The resulting AP configuration is specific to the floor layout, candidate installation sites, AP specifications, user-demand assumptions, and installation costs used in the model.

Recent WLAN planning research formulates indoor AP deployment as a binary linear program that jointly selects AP count, candidate locations, and emission channels to satisfy coverage requirements while controlling infrastructure cost [21]. Joint AP placement and channel assignment can also be expressed as an integer linear program that incorporates power tuning and partially overlapping channels to manage interference and avoid costly post-deployment adjustments [29]. Wireless Internet access planning on campuses also fits classic facility-location structure, with maximal covering models used to reconfigure access point sites to improve service quality and reduce operating costs [30].

Related router siting problems in mesh networks use mixed-integer linear programming variants of the p-median model to minimize the number of deployed routers under coverage-radius constraints [31]. Cellular network design similarly applies mixed-integer programming to choose base-station locations that minimize installation and operating costs under capacity and QoS constraints while computing path loss via propagation modeling [32]. These recent formulations support integer programming as a defensible core method for selecting AP locations under explicit coverage, cost, and propagation constraints [21].

Existing studies have examined AP placement through integer programming, genetic algorithms, particle swarm optimization, and hybrid methods under various coverage and configuration constraints. The present study focuses more narrowly on a site-specific minimum-cost formulation for a bounded university floor, using explicit coverage, bandwidth, range, and AP-count constraints.

A. Basic Concepts

- Wi-Fi refers to a family of wireless local area network (WLAN) technologies defined by the IEEE 802.11 standards. It enables devices such as desktop and laptop computers, smartphones, wearables, printers, and cameras to connect to a local network and access the Internet without physical cabling, typically within a limited coverage area.
- Wide area network (WAN) is a telecommunications network that links servers and end-user devices across large geographic distances, ranging from hundreds to thousands of kilometers, often by interconnecting multiple local or metropolitan networks.
- Local area network (LAN) is a telecommunications network that connects servers and end-user devices within a limited geographic area, typically within a room, building, or campus, and usually within about one kilometer.
- Access point (AP) is a network device that provides Wi-Fi connectivity, allowing wireless clients to join a wired network. In many deployments, the AP connects by Ethernet to upstream equipment such as a switch and router, which then provides access to local network services and the Internet. Some products combine AP and router functions in a single unit, and some broadband gateways (modem-router devices) also include built-in Wi-Fi, effectively serving as the access point. In larger organizations, multiple APs are installed across the site to extend coverage so users can connect to the same network from different areas.
- Routers receive and transmit data through computer networks. By sending data packets to their intended IP addresses, it manages traffic between different networks and permits several devices to share an Internet connection.
- Bandwidth is the maximum data speed that a network connection can support. Greater bandwidth suggests a higher capacity, but it doesn't always translate into better performance.

II. Objectives of the Study

The study aims to determine an access point configuration for the third floor of the College of Technologies that satisfies coverage and bandwidth requirements while minimizing installation cost under defined constraints. Specifically, it seeks to:

1) Assess the existing AP locations of a given network in terms of:

a) Coverage quality (minimal path loss)

b) Model type (number of users, bandwidth (2.4 or 5G, area of coverage), Environment type (indoor or outdoor), Obstacle type, and

c) Cost metrics (installation cost per unit, fiber optic cable length cost).

2) Design an integer programming model to identify optimal AP locations, specifically addressing:

a) Coverage satisfaction

b) Cost minimization

c) Bandwidth feasibility

III. Methodology

This study is based on the existing access point (AP) locations on the third floor of the College of Technologies (COT) at Bukidnon State University, as reflected in Fig. 1.

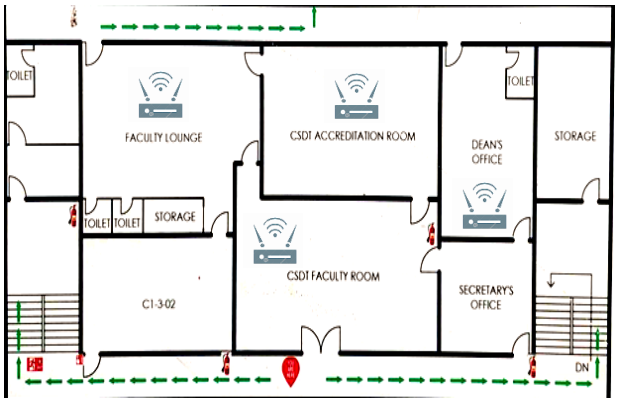


Fig. 1. Floor plan of the COT 3rd floor.

Fig. 1 presents the third-floor layout of the College of Technologies, which served as the spatial basis for identifying candidate AP locations, service areas, and AP-to-area distances

used in the optimization model. Because the floor layout, AP specifications, demand assumptions, and installation costs are site-specific, the resulting configuration should also be interpreted as site-specific.

Fig. 2 shows the existing network topology of the COT 3rd floor, indicating the current access points and the offices they serve within the study site. Together, Fig. 1 and Fig. 2 provide the spatial and network context used in formulating the optimization model.

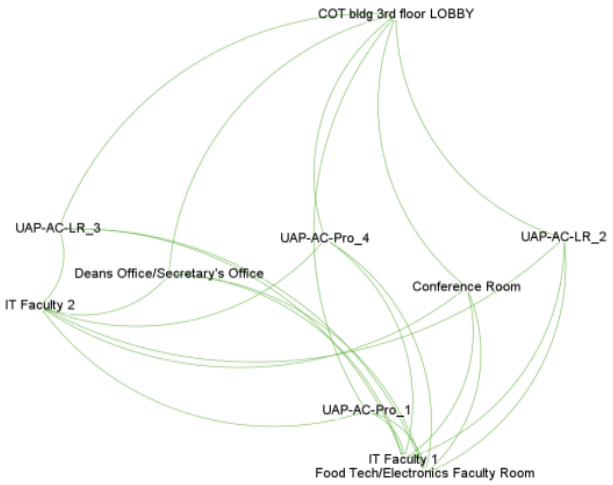


Fig. 2. COT 3rd floor offices served by 4 APs.

Table I provides an overview of four APs at the COT, detailing their locations, user coverage, specifications, unit costs, and cable distances from the switch used in the computation of installation cost. The network infrastructure utilizes Cat 6 cables, priced at Php 35 per meter, to support data transmission rates of up to 10 Gbps from the switch, with 1 Gbps allocated per access point (AP). Each AP offers dual-band operation with maximum data rates of 450 Mbps on the 2.4 GHz frequency and 867 Mbps on the 5 GHz frequency. Faculty, personnel, and students are limited to 60 Mbps, 40 Mbps, and 20 Mbps per user per device, respectively, with each AP capable of handling up to 175 connections simultaneously (70% of 250).

A. Optimization Method

This study formulates the access point placement problem as a binary integer programming model. The model determines which access points should be installed and which service areas should be assigned to the installed access points, subject to coverage, bandwidth, range, and installation limits. The objective is to minimize the total installation cost while ensuring that all required service areas are served within the defined constraints.

Let $i \in I$ denote the set of candidate access points and $j \in J$ denote the set of service areas.

B. Objective Function

The objective of the model is to minimize the total installation cost of the selected access points:

$$\text{Minimize } z = \sum_{i \in I} C_i x_i \quad (1)$$

1) Decision Variables

$$X_i = \begin{cases} 1, \text{ if access point } i \text{ is installed} \\ 0, \text{ otherwise} \end{cases}$$

$$Y_{ij} = \begin{cases} 1, \text{ if area } j \text{ is served by access point } i \\ 0, \text{ otherwise} \end{cases}$$

2) Parameters

C_i : total installation cost of access point i, consisting of the unit cost of the AP and the cable cost from the switch to the AP location

r_i : coverage range of access point i

$d_{i,j}$: distance between access point i and area j

B_i : maximum bandwidth capacity of access point i

U_j : total bandwidth demand of users in area j

L : maximum number of access points allowed for installation

$$a_{ij} = \begin{cases} 1, \text{ if } d_{ij} \le r_i \\ 0, \text{ otherwise} \end{cases}$$

where,

a_{ij} indicates whether service area j lies within the allowable range of access point i.

The attenuation-adjusted broadcast value for each AP-to-area pair was computed by combining the nominal broadcast value of the access point with an attenuation term based on distance and wall barriers:

$$P_{ij}^{(2.4)} = B_i^{(2.4)} + \left[L_0^{(2.4)} + 10n \log_{10}\left(\frac{d_{ij}}{d_0}\right) + N_{iJ} W_{2.4} \right] \quad (2)$$

$$P_{ij}^{(5)} = B_i^{(5)} + \left[L_0^{(5)} + 10n \log_{10}\left(\frac{d_{ij}}{d_0}\right) + N_{iJ} W_5 \right] \quad (3)$$

TABLE I. EXISTING AP LOCATIONS

AP	AP Current Location	Network Connected	AP Model/Type	Transmission Frequency	Range	Environment Type	Cost of AP	Distance of AP from switch(m)
1	COT 3rd Floor Lobby	Student, Faculty, Personnel	UAP-AC-Pro	5 GHz and 2.4GHz	up to 122m	Indoor	Php 12,100.00	20.92
2	COT Conference Room	Student, Faculty	UAP-AC-Lite	5 GHz and 2.4GHz	up to 91 m	Indoor	Php 9,000.00	43.6
3	COT Food Tech Faculty	Student, Faculty, Personnel	UAP-AC-LR	5 GHz and 2.4GHz	up to 183 m	Indoor	Php 14,000.00	32.75
4	COT IT Faculty	Faculty	UAP-AC-LR	5 GHz and 2.4GHz	up to 183 m	Indoor	Php 14,000.00	24.49

where, $B_i^{(2..4)}$ and $B_i^{(5)}$ are the nominal broadcast values of access point i for the 2.4 GHz and 5 GHz bands, $L_0^{(2.4)}$ and $L_0^{(5)}$ are the reference-loss constants, n is the path-loss exponent, d_{ij} is the distance between access point *i* and service area *j*, d_0 is the reference distance, $N_{i,j}$ is the number of wall barriers between access point *i* and service area *j*, and $W_{2.4}$ and W_5 are the attenuation values assigned per wall barrier for the 2.4 GHz and 5 GHz bands, respectively.

The attenuation-related calculations followed standard indoor radio-planning principles, with Recommendation ITU-R P.1238 included as a relevant reference for indoor radiocommunication system planning [28]. The propagation analysis incorporated both distance-related attenuation and barrier-related attenuation between each access point and service area. The resulting attenuation-adjusted values were used as supporting indicators in interpreting AP-to-area relationships within the study site.

The total bandwidth demand for each service area was estimated from the number of students, faculty members, and personnel assigned to the area, their corresponding per-user bandwidth allowances, and a simultaneous-use factor representing the proportion of users expected to be active on the network at the same time. Thus:

$$U_j = \tau_j(20S_j + 60F_j + 40P_j) \quad (4)$$

where, S_j, F_j, and P_j represent the number of students, faculty members, and personnel in service area *j*, respectively; and τ_j denotes the simultaneous-use factor for area *j*.

The simultaneous-use factor τ_j represents the proportion of users in service area *j* expected to be active on the network at the same time. This factor was used to estimate concurrent service demand rather than assuming that all assigned users would simultaneously consume their full bandwidth allowance. In this study, τ_j was assigned according to the expected activity level of each area, with lower values used for transitional or administrative spaces and higher values used for high-concurrency spaces such as the Conference Room. This approach is consistent with WLAN capacity-planning guidance that bases demand on expected active users and connections rather than total assigned occupancy.

For the present study, the simultaneous-use factor was set at 0.30 for the lobby, 0.80 for the Conference Room, 0.50 for faculty rooms, and 0.40 for the Dean's Office/Secretary's Office.

The total installation cost for each access point was computed as the sum of the AP unit cost and the cable cost from the switch to the AP location. Thus:

$$C_i = A_i + 35L_i \quad (5)$$

where, A_i is the unit cost of access point *i*, L_i is the cable length in meters from the switch to the AP location, and 35 represents the cable cost per meter in Philippine pesos at the time of the conduct of this study.

3) Constraints: The following constraints are imposed in minimizing access point placement.

a) Coverage constraint: Each area j must be assigned to at least one access point within its range:

$$\sum_{i\in I} Y_{ij} \geq 1, \forall j \in J \quad (6)$$

b) AP selection constraint: An area j can only be served by access point i if access point i is installed:

$$Y_{ij} \leq X_i, \forall i \in I, \forall j \in J \quad (7)$$

c) Bandwidth constraint: The total bandwidth demand assigned to access point i must not exceed its bandwidth capacity:

$$\sum_{(j\in J)} U_j Y_{ij} \leq B_i X_i, \forall i \in I \quad (8)$$

d) Range constraint: Access point i can only serve area j if area j lies within its coverage range:

$$Y_{ij} \leq a_{ij}, \forall i \in I, \forall j \in J \quad (9)$$

e) AP count constraint: The total number of installed access points must not exceed the allowable limit:

$$\sum_{i\in I} X_i \leq L \quad (10)$$

f) Binary restriction:

$$X_i \in \{0,1\}, \; Y_{ij} \in \{0,1\} \quad (11)$$

This formulation provides the basis for selecting access point locations that satisfy the required service coverage and bandwidth demand while minimizing total installation cost within the defined study constraints.

C. Solution Procedure

The integer programming model was implemented in Microsoft Excel Solver. The objective cell was set to minimize the total installation cost, and the changing variable cells corresponded to the binary decision variables for AP installation and AP-to-area assignment. The model was solved using the Simplex LP method with binary restrictions imposed on X_i and Y_{ij}.

IV. Results

Under the formulated binary integer programming model and the specified Excel Solver settings, Microsoft Excel Solver produced a feasible minimum-cost AP assignment for the evaluated service areas of the COT third floor. The final Solver run required 2.35 seconds, and the resulting solution was accepted in this study as the optimal solution for the study instance. The evaluated service areas included the COT third-floor lobby, Conference Room, Food Tech/Electronics Faculty Room, IT Faculty 1, IT Faculty 2, and the Dean's Office/Secretary's Office.

The final solution assigns binary values to the decision variables X_i and Y_{ij}, consistent with the binary integer programming formulation of the study. In this model, X_i=1 indicates that access point i is selected in the final solution, whereas X_i=0 indicates that it is not selected. Similarly, Y_{ij}=1 indicates that service area j is assigned to access point i, whereas Y_{ij}=0 indicates that no such assignment is made. These values were generated by the Solver and form the basis for interpreting the final AP-selection and AP-assignment pattern.

Table II presents the evaluated AP-to-area assignments under the optimization model. The table shows that all evaluated AP-area pairs satisfy the coverage-feasibility condition. This indicates that, for the present study site, all candidate access points are within the allowable range of the evaluated service areas under the specified model assumptions. As a result, the range condition operates as a general feasibility requirement rather than as the principal factor distinguishing selected from unselected AP-area pairings.

Because all candidate AP-area relationships are feasible in terms of range, the differentiating output of the model is reflected in the Solver-generated binary variables X_i and Y_{ij}. The values of these variables identify which access points were retained in the final solution and which AP-area assignments were selected after the cost, bandwidth, and AP-count constraints were applied. In this sense, Table II reports not only physical feasibility but also the discrete decisions made by the optimization model under the imposed constraints.

TABLE II. EVALUATION OF AP-TO-AREA ASSIGNMENTS UNDER THE OPTIMIZATION MODEL

Area	Access Point AP_i	Distance between AP i and area j (m) $d_{i,j}$		Maximum data rate broadcast B_i		Attenuation-adjusted broadcast values P_{ij}		Total cost per AP c_i	AP range (m) r_i	Total bandwidth demand of users in area j u_j	x_i	y_{ij}	Coverage feasibility $d_{ij} <= r_i$
				2.4 Ghz	5Ghz	2.4 Ghz	5Ghz						
COT bldg 3rd floor	UAP-AC-Pro	d1,1	8	450	1300	450.00	1300.00	13,007.20	122	180	1	1	1
	UAP-AC-Lite	d2,1	15	300	867	300.00	867.00	9,907.20	91	180	1	1	1
	UAP-AC-LR	d3,1	6	450	867	450.00	867.00	14,907.20	183	180	0	0	1
	UAP-AC-LR	d4,1	10.4	450	867	450.00	867.00	14,907.20	183	180	0	0	1
Conference Room	UAP-AC-Pro	d1,2	11	450	1300	450	1300	13,801.00	122	1160	0	0	1
	UAP-AC-Lite	d2,2	3.4	300	867	300	867	10,701.00	91	1160	0	0	1
	UAP-AC-LR	d3,2	9	450	867	450	867	15,701.00	183	1160	0	0	1
	UAP-AC-LR	d4,2	10	450	867	450	867	15,701.00	183	1160	0	0	1
Food Tech/Electronics Faculty Room	UAP-AC-Pro	d1,3	7	450	1300	450	1300	13,421.25	122	474	0	0	1
	UAP-AC-Lite	d2,3	10	300	867	300	867	10,321.25	91	474	0	0	1
	UAP-AC-LR	d3,3	3.4	450	867	450	867	15,321.25	183	474	0	0	1
	UAP-AC-LR	d4,3	9	450	867	450	867	15,321.25	183	474	0	0	1
IT Faculty 1	UAP-AC-Pro	d1,4	17	450	1300	450	1300	13,132.15	122	420	1	1	1
	UAP-AC-Lite	d2,4	13	300	867	300	867	10,032.15	91	420	0	0	1
	UAP-AC-LR	d3,4	14	450	867	450	867	15,032.15	183	420	0	0	1
	UAP-AC-LR	d4,4	2.8	450	867	450	867	15,032.15	183	420	0	0	1
IT Faculty 2	UAP-AC-Pro	d1,5	15	450	1300	450	1300	13,272.50	122	294	0	0	1
	UAP-AC-Lite	d2,5	15.2	300	867	300	867	10,172.50	91	294	0	0	1
	UAP-AC-LR	d3,5	11	450	867	450	867	15,172.50	183	294	0	0	1
	UAP-AC-LR	d4,5	5	450	867	450	867	15,172.50	183	294	0	0	1
Deans Office/Secretary's Office	UAP-AC-Pro	d1,6	10	450	1300	450	1300	13,850.00	122	104	0	0	1
	UAP-AC-Lite	d2,6	5	300	867	300.00	867.00	10,750.00	91	104	0	0	1
	UAP-AC-LR	d3,6	11	450	867	450.00	867.00	15,750.00	183	104	0	0	1
	UAP-AC-LR	d4,6	16.6	450	867	450.00	867.00	15,750.00	183	104	0	0	1

V. DISCUSSION

The results show that the final AP configuration must be interpreted mainly through the Solver-generated binary variables X_i and Y_{ij}. Since all evaluated AP-area pairs satisfy the coverage-feasibility condition, the selected pattern of X_i and Y_{ij} values should be interpreted as the result of constrained optimization among feasible alternatives rather than as a consequence of range exclusion.

Table II shows that all evaluated AP-to-area pairs satisfy the coverage-feasibility condition. This indicates that the candidate access points are within the allowable range of the evaluated service areas under the assumptions of the study. Under this condition, range functions mainly as a feasibility requirement and does not strongly differentiate among candidate assignments. The final configuration should therefore be interpreted as a constrained selection among feasible alternatives rather than as an outcome driven primarily by distance exclusion.

The selected pattern of X_i and Y_{ij} values further indicates that the optimization process filtered feasible options and retained only those selections and assignments that satisfied the imposed constraints at minimum total installation cost. In this way, a value of 1 reflects a selected decision in the final solution, while a value of 0 reflects a feasible option that was not retained by the Solver under the defined optimization criteria.

Because the nominal AP ranges are large relative to the dimensions of the study site, installation cost and bandwidth feasibility exert a stronger influence on the final AP assignment pattern than range limitation alone. This supports the interpretation that the binary integer programming model is useful as a structured planning tool for identifying a feasible minimum-cost AP configuration in a bounded indoor environment. At the same time, the findings should be read within the limits of the present formulation and site conditions.

The present study applies the binary integer programming model under a single set of site-specific assumptions for user demand, AP range, and installation cost. Accordingly, the resulting configuration should be interpreted as the optimal solution for the defined study instance rather than as a robustness-tested solution across multiple scenarios. This study does not include scenario-based analysis under varying user density, coverage-radius assumptions, or budget constraints. As a result, the discussion of model sensitivity is limited to literature-supported expectations and to the recognition that different parameter settings may produce different feasible configurations in future applications of the model.

VI. CONCLUSION

The study shows that binary integer programming can be applied as a structured approach to access point placement on the third floor of the College of Technologies at Bukidnon State University. Under the defined assumptions and constraints, the model identified a feasible minimum-cost AP configuration that satisfied the required coverage and bandwidth conditions of the study area.

The results indicate that all evaluated AP-to-area pairs were range-feasible under the specified model assumptions. Consequently, the final configuration was influenced more strongly by installation cost and bandwidth feasibility than by range limitation alone. This highlights the value of binary integer programming as a method for selecting among multiple feasible alternatives under explicit operational constraints.

Because the study is based on site-specific physical, technical, and cost parameters, the resulting AP configuration should be interpreted as specific to the modeled environment. The same formulation may still be applied to other indoor settings, provided that the relevant floor layout, access point specifications, demand assumptions, and installation costs are recalibrated to reflect the target context. A post-deployment wireless site survey is recommended to verify actual field performance and identify any remaining coverage gaps.

ACKNOWLEDGMENT

The researchers express their profound gratitude to the Bukidnon State University for generously providing the resources, facilities, and support essential for the successful completion of this study. Special appreciation is extended to the Center for Mathematical Innovations for their invaluable guidance and technical assistance throughout the duration of the project.

REFERENCES

[1] D. Chizhik, J. Du, and R. A. Valenzuela, "Universal path gain laws for common wireless communication environments," IEEE Trans. Antennas Propag., vol. 70, no. 4, pp. 2928–2941, 2022, doi: 10.1109/TAP.2021.3121173.

[2] G. N. T. Cagande, J. L. Abella, J. L. Q. Rojo, and K. J. R. Caseres, "Network design for Bukidnon State University College of Technologies Annex B Campus," Int. J. Eng. Trends Technol., vol. 72, no. 11, pp. 207–224, 2024, doi: 10.14445/22315381/IJETT-V72I11P122.

[3] N. R. Leonor, S. Faria, G. L. Ramos, P. V. Gonzalez Castellanos, C. Rodríguez, L. da S. Mello, and R. F. S. Caldeirinha, "Site-specific radio propagation model for macrocell coverage at sub-6 GHz frequencies," IEEE Trans. Antennas Propag., vol. 70, no. 10, pp. 9706–9715, 2022, doi: 10.1109/TAP.2022.3184416.

[4] A. Mehina and I. Čavlina, "Signal propagation," Nauka i Tehnologija, vol. 11, no. 1, pp. 47–54, 2023, doi: 10.58952/nit20231101047.

[5] J. C. da Silva and E. Costa, "A ray-tracing model for millimeter-wave radio propagation in dense-scatter outdoor environments," IEEE Trans. Antennas Propag., vol. 69, no. 12, pp. 8618–8629, 2021, doi: 10.1109/TAP.2021.3083779.

[6] G. Kara and C. Özturan, "Parallel network simplex algorithm for the minimum cost flow problem," Concurrency Computat. Pract. Exper., vol. 34, no. 4, Art. no. e6659, 2022, doi: 10.1002/cpe.6659.

[7] B. Golden, L. Schrage, D. Shier, and L. A. Apergi, "The unexpected power of linear programming: An updated collection of surprising applications," Ann. Oper. Res., vol. 343, no. 2, pp. 573–605, 2024, doi: 10.1007/s10479-024-06245-5.

[8] A. B. Beyi and T. A. Godana, "The maximum attainable flow and minimal cost problem in a network," Amer. J. Math. Comput. Modell., vol. 9, no. 2, pp. 22–37, 2024, doi: 10.11648/j.ajmcm.20240902.1.

[9] I. Joseph, E. Oghu, and O. O. Roberts, "Path loss and models: A survey and future perspective for wireless communication networks," Int. J. Adv. Netw. Appl., vol. 15, no. 2, pp. 5892–5907, 2023, doi: 10.35444/IJANA.2023.15209.

[10] S. Cheerla, "Analysis of different path loss models in urban suburban and rural environment," Int. J. Emerg. Trends Eng. Res., vol. 8, no. 7, pp. 2972–2976, 2020, doi: 10.30534/ijeter/2020/14872020.

[11] H. P. Keeler, B. Błaszczyszyn, and E. Cali, "Connectivity and interference in device-to-device networks in Poisson-Voronoi cities," in Proc. 21st Int. Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2023, pp. 1–8, doi: 10.23919/WiOpt58741.2023.10349879.

[12] N. A. M. Maung and W. Zaw, "Comparative study of RSS-based indoor positioning techniques on two different Wi-Fi frequency bands," in Proc. 17th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON), 2020, pp. 185–188, doi: 10.1109/ECTI-CON49241.2020.9158211.

[13] M. M. Ahamed and S. Faruque, "5G network coverage planning and analysis of the deployment challenges," Sensors, vol. 21, no. 19, Art. no. 6608, 2021, doi: 10.3390/s21196608.

[14] M. S. Haroon, Z. H. Abbas, F. Muhammad, and G. Abbas, "Analysis of coverage-oriented small base station deployment in heterogeneous cellular networks," Phys. Commun., vol. 38, Art. no. 100908, 2020, doi: 10.1016/j.phycom.2019.100908.

[15] M. Pagin, T. Zugno, M. Polese, and M. Zorzi, "Resource management for 5G NR integrated access and backhaul: A semi-centralized approach," IEEE Trans. Wireless Commun., 2021, doi: 10.1109/TWC.2021.3098967.

[16] A. Fayad, T. Cinkler, and J. Rak, "5G/6G optical fronthaul modeling: Cost and energy consumption assessment," J. Opt. Commun. Netw., vol. 15, no. 9, p. D33, 2023, doi: 10.1364/JOCN.486547.

[17] R. Ramesh, M. Arunachalam, H. K. Atluri, C. K. S., et al., "LoRaWAN for smart cities: Experimental study in a campus deployment," in LPWAN Technologies for IoT and M2M Applications, pp. 327–345, 2020, doi: 10.1016/B978-0-12-818880-4.00016-8.

[18] R. Akram, A. H. Al-Nakkash, O. N. M. Salim, and A. A. S. AlAbdullah, "Proposed APs distribution optimization algorithm: Indoor coverage solution," J. Phys.: Conf. Ser., vol. 1804, no. 1, Art. no. 012134, 2021, doi: 10.1088/1742-6596/1804/1/012134.

[19] W. Nur, A. Aziz, Marianah Masrie, R. Sam, and Mohd., "WiFi approximated strength measurement method with brute force algorithm for a minimum number of AP and maximum WiFi coverage," in Proc. IEEE 10th Symp. Comput. Appl. Ind. Electron. (ISCAIE), 2020, pp. 180–185, doi: 10.1109/ISCAIE47305.2020.9108833.

[20] P. Liu, Q. Hu, K. Jin, G. Yu, and Z. Tang, "Toward the energy-saving optimization of WLAN deployment in real 3-D environment: A hybrid swarm intelligent method," IEEE Syst. J., vol. 16, no. 2, pp. 2425–2436, 2022, doi: 10.1109/JSYST.2021.3065434.

[21] A. Oustry, M. Le Tilly, T. H. Clausen, C. D'Ambrosio, and L. Liberti, "Optimal deployment of indoor wireless local area networks," Networks, vol. 81, no. 1, pp. 23–50, 2023, doi: 10.1002/net.22116.

[22] M. Sun, Y. Wang, N. Zheng, Q. Wang, G. Chen, and Z. Li, "Enhanced genetic algorithm-based Wi-Fi access points deployment for RTT positioning: Fitness function design and analysis," CEUR Workshop Proc., vol. 3919, 2024.

[23] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: Past, present, and future," Multimed. Tools Appl., vol. 80, no. 5, pp. 8091–8126, 2021, doi: 10.1007/s11042-020-10139-6.

[24] O. S. Naif and I. J. Mohammed, "Wireless optimization algorithm for indoor placement (WOAIP)," J. Phys.: Conf. Ser., vol. 1963, no. 1, Art. no. 012028, 2021, doi: 10.1088/1742-6596/1963/1/012028.

[25] A. G. Gad, "Particle swarm optimization algorithm and its applications: A systematic review," Arch. Comput. Methods Eng., vol. 29, no. 5, pp. 2531–2561, 2022, doi: 10.1007/s11831-021-09694-4.

[26] F. A. Ovat and A. A. Josephat, "The particle swarm optimization (PSO) algorithm application: A review," Global J. Eng. Technol. Adv., vol. 3, no. 3, pp. 1–6, 2020, doi: 10.30574/gjeta.2020.3.3.0033.

[27] R. G. Garroppo, G. Nencioni, L. Tavanti, B. Gendron, and M. G. Scutellà, "Optimal access point power management for green IEEE 802.11 networks," Sensors, vol. 21, no. 6, Art. no. 2076, 2021, doi: 10.3390/s21062076.

[28] ITU-R, Recommendation ITU-R P.1238-13: Propagation data and prediction methods for the planning of indoor radiocommunication systems and radio local area networks in the frequency range from 300 MHz to 450 GHz, 2025.

[29] B. P. Tewari and S. C. Ghosh, "Efficient AP placement through power control and partially overlapping channel assignment," Wireless Pers. Commun., vol. 112, pp. 1221–1241, 2020, doi: 10.1007/s11277-019-06721-2.

[30] B. A. Pludow, A. T. Murray, and R. L. Church, "Service quality modeling to support optimizing facility location in a microscale environment," Socioecon. Plann. Sci., vol. 82, Art. no. 101273, 2022, doi: 10.1016/j.seps.2022.101273.

[31] M. F. Mandarino, T. A. P. Beneteli, and L. P. Cota, "A decision support system based on a mixed-integer linear programming model for location of routers in open-pit mines," in Proc. 26th Int. Conf. Enterprise Inf. Syst. (ICEIS), vol. 1, 2024, pp. 620–627, doi: 10.5220/0012586400003690.

[32] Y. Almoghathawi, H. Bin Obaid, and S. Selim, "Optimal location of base stations for cellular mobile network considering changes in users locations," J. Eng. Res., vol. 13, no. 2, pp. 561–567, 2025, doi: 10.1016/j.jer.2024.04.020.

AI and Blockchain for Secure Healthcare Data Management: A Bibliometric Analysis of Research Trends and Thematic Clusters (2020–2025)

Maroufi Mohammed, Lamzabi Siham, Ziti Soumia
Intelligent Processing and Security of Systems-Faculty of Sciences, Mohammed V University in Rabat, Morocco

***Abstract*—The convergence of artificial intelligence (AI) and blockchain has become an active axis of interdisciplinary research in healthcare data security. This paper reports a bibliometric analysis of 434 Scopus-indexed articles published between 2020 and 2025, with data collection and processing performed on 20 April 2026. The objective is to map the intellectual structure, collaborative dynamics, and thematic composition of this expanding field. The corpus was analyzed using the bibliometrix R package (version 4.3.0) and VOSviewer (version 1.6.20). The increasing research output is evidenced by an annual growth rate of 34.8%, with publication volume growing from 10 articles in 2020 to a maximum of 152 articles in 2025 across a total of 198 unique venues. Keyword co-occurrence analysis, processed through the Louvain community detection algorithm on 250 high-frequency terms with association-strength normalization, produced four thematic clusters: Blockchain and Privacy-Preserving Techniques, Healthcare Systems and Cybersecurity Infrastructure, Artificial Intelligence and Clinical Diagnostics, and Electronic Health Records and Interoperability. India, China, Saudi Arabia, and the United States lead scholarly output. The international co-authorship rate of 49.31% reflects the globally distributed nature of the research community. IEEE Access and the IEEE Journal of Biomedical and Health Informatics are the dominant publication venues. Federated learning occupies a structurally central position, with a betweenness centrality of 101.5, acting as the principal methodological bridge between the two technologies. An average of 27.96 citations per document confirms the above-average scholarly impact of the corpus. The results provide researchers, practitioners, and policymakers with an evidence-based map of the field's trajectory, its most productive research directions, and its remaining structural gaps. The underlying dataset, search logs, and analysis scripts are released openly to support full reproducibility.**

Keywords*—*Artificial intelligence; blockchain; healthcare data security; bibliometric analysis; federated learning; internet of medical things; data privacy; cybersecurity; science mapping; VOSviewer

I. Introduction

Healthcare systems now produce data at a scale that was difficult to imagine even a decade ago. Electronic health records (EHR), wearable sensors, mobile health applications, and Internet of Medical Things (IoMT) devices generate continuous streams of clinical, physiological, and behavioral information across hospitals, clinics, and patient homes. The rapid accumulation of this data has far-reaching implications for diagnosis, treatment personalization, and population health monitoring, but it has also exposed healthcare institutions to security risks that legacy centralized infrastructures were never designed to address [1]. Ransomware campaigns targeting hospitals, unauthorized access incidents affecting patient records, and data integrity breaches in connected medical devices have become recurring events in the post-pandemic period [2].

Two technologies have received sustained scholarly attention as candidate responses to this situation. Artificial intelligence, which groups machine learning, deep learning, and the more recent paradigm of federated learning, is used to detect anomalies, identify cyber threats, and automate decisions at scales that human analysts cannot match [3]. Blockchain, with its decentralized and immutable ledger structure, provides verifiable data provenance, tamper-resistant storage, and programmable access through smart contracts [4]. The combination of the two is increasingly presented as a comprehensive strategy for securing healthcare data flows, particularly in IoMT environments where data integrity must be preserved across heterogeneous devices and organizational boundaries [5].

Despite the growing body of primary research at this intersection, the field lacks a comprehensive mapping of its intellectual structure. Earlier reviews have generally focused on one technology at a time, on narrow application domains, or on qualitative synthesis without bibliometric grounding. Few studies have quantified the geographic distribution, institutional concentration, or thematic organization of scholarly production, leaving researchers without a reliable overview of the field's maturity and gaps. The absence of such a cartography is significant given the pace of publication growth in this area and the emergence of federated learning as a dominant methodological paradigm from 2021 onward [6].

This study addresses the gap through a structured bibliometric analysis of Scopus-indexed literature on AI-blockchain integration for healthcare data security over the period 2020–2025, with data collection and processing performed on 20 April 2026. Bibliometric analysis is well-suited to this objective. In contrast to narrative reviews, it allows the quantitative examination of publication trends, citation networks, authorship patterns, and thematic clustering, producing reproducible evidence-based cartographies of a scientific field [7]. This approach uses the Louvain community detection algorithm on co-occurrence matrices which is able to identify research communities that would otherwise go unnoticed in traditional review methods [8].

The analysis is guided by three research questions.

RQ1 concerns the temporal and geographic trends in scholarly production on AI and blockchain for healthcare data security over the 2020–2025 period.

RQ2 investigates which journals, authors, and institutions form the most influential nodes in this research network.

RQ3 examines the thematic clusters that structure the conceptual landscape of the field and the central research threads within each cluster.

The study contributes to the literature in three respects. The first contribution is a quantitative treatment of the AI-blockchain-healthcare nexus based on a corpus of 434 documents from 198 venues, which is broader than any prior bibliometric study focused on this intersection. The second is the combined use of bibliometrix and VOSviewer, which produces both statistical indicators and visually interpretable network maps. The third is the identification of three structural gaps in the field: the underdevelopment of interoperability research relative to its practical importance, the limited attention paid to regulatory and ethical dimensions, and the concentration of research output in a small number of countries.

The remainder of the paper is organized as follows. Section II reviews earlier bibliometric and narrative studies on related topics and specifies the research gap addressed here. Section III describes the data sources, search query, analytical methods, and reproducibility protocol. Section IV reports and discusses the results across seven analytical dimensions. Section V concludes with research implications and directions for future work.

II. Related Works

Prior bibliometric and review studies have examined aspects of the AI–blockchain–healthcare nexus, though none have produced a comprehensive mapping of the intersection focused specifically on data security and privacy. This section reviews the principal prior contributions, positions the present study with respect to each, and specifies the gap it fills.

A. Bibliometric and Narrative Studies on Blockchain in Healthcare

In an early science mapping study, Agbo et al. [9] surveyed blockchain applications in healthcare and identified data sharing and EHR management as dominant themes in the pre-2019 literature. Their analysis predated the widespread adoption of federated learning and therefore did not capture the role that privacy-preserving machine learning has come to play in shaping the field. Treatment of AI in that work remained peripheral rather than structural. Hasselgren et al. [15] addressed a similar scope through a scoping review that used co-citation analysis to identify intellectual foundations of blockchain in medical records management. Their temporal window ended before the acceleration of federated learning research in 2021–2022, and AI-related publications were not part of their analytical scope. More recently, Casino et al. [16] provided a systematic literature review of blockchain-based applications across industries including healthcare. Their taxonomy is useful for locating healthcare within the wider blockchain landscape, but the breadth of their scope diluted the specificity of healthcare-related findings, and privacy-preserving AI did not feature as an organizing theme.

B. Bibliometric Studies on AI in Healthcare

Secinaro et al. [10] conducted a structured bibliometric review of AI in healthcare, documenting the rapid growth of deep learning applications in diagnostics and treatment optimization. Their work is useful for characterizing the trajectory of AI research in clinical settings but did not address the security dimension and did not treat blockchain as a structural element. Javaid et al. [11] reviewed machine learning applications in healthcare in a similar register, without integrating blockchain, which reproduces the conceptual separation that the present work seeks to overcome.

C. Narrative Treatments of the AI–Blockchain Combination

A limited number of studies have attempted to combine treatments of AI and blockchain in healthcare. Bathula et al. [12] offered a narrative review describing what they called the AI–blockchain–healthcare tripod, arguing that the three elements jointly support secure and scalable medical systems. Their approach was qualitative and did not employ bibliometric techniques, which left open questions about citation impact, geographic concentration, and thematic evolution. Omidian [13] and Alshehri [14] provided descriptive surveys that identified similar themes but lacked the rigor of a formal science mapping methodology.

D. Foundational References on Federated Learning

Recent contributions have addressed the AI–blockchain–security nexus more directly through the lens of federated learning. Banabilah et al. [17] published a comprehensive review of federated learning that has become a central reference in the field, with 531 citations recorded in April 2026. Their work establishes federated learning as the methodological centerpiece of privacy-preserving AI but does not provide a bibliometric mapping of the surrounding research field. Ferrag et al. [18] offered an experimental analysis of federated deep learning for IoT cybersecurity that has accumulated 301 citations and is regularly referenced in healthcare-specific work. Tanwar et al. [19] examined the adoption of machine learning in blockchain-based applications and reached 335 citations, becoming a foundational reference for subsequent integration studies.

E. Research Gap and Positioning

Table I summarizes the scope, methodology, and limitations of the most relevant prior studies and positions the present contribution against them.

The present study extends this literature in three respects. First, it applies a complete bibliometric protocol, including growth analysis, Bradford's law, co-authorship network analysis, and Louvain-based keyword clustering, to a corpus of 434 documents with an explicit security and privacy focus. Second, the 2020–2025 temporal window captures the full arc of a field that experienced its steepest growth during this period, and the April 2026 data collection incorporates the latest available citation signals. Third, the combined use of bibliometrix and VOSviewer produces both quantitative indicators and network visualizations, an approach

recommended by Donthu et al. [20] for rigorous science mapping.

TABLE I. POSITIONING WITH RESPECT TO PRIOR BIBLIOMETRIC AND REVIEW STUDIES

Study	Scope	Method	AI treated?	Security focus?	Period covered
Agbo et al. [9]	Blockchain in healthcare	Systematic review	Peripheral	No	Pre-2019
Hasselgren et al. [15]	Blockchain in medical records	Scoping review + co-citation	No	Partial	Pre-2020
Casino et al. [16]	Blockchain across industries	Systematic literature review	No	No	Pre-2019
Secinaro et al. [10]	AI in healthcare	Structured literature review	Yes	No	Pre-2021
Javaid et al. [11]	ML in healthcare	Narrative review	Yes	No	Pre-2022
Bathula et al. [12]	AI–blockchain–healthcare	Narrative review	Yes	Partial	Pre-2024
Present study	AI–blockchain–healthcare security	Full bibliometric protocol	Yes	Yes	2020-2025

III. RESEARCH METHODOLOGY

A. Data Source and Query

The Scopus database was chosen as the exclusive data source for this analysis. Scopus offers broad interdisciplinary coverage, with particular depth in computer science, engineering, and biomedical sciences, which are the three domains most directly relevant to this study. Its structured export format is compatible with the bibliometrix package, which supports reproducibility [7]. Baas et al. [21] and Mongeon and Paul-Hus [22] have documented the comparative strengths of Scopus in interdisciplinary settings, noting its wider journal coverage in engineering and health informatics compared to Web of Science. The exclusive use of Scopus is acknowledged as a limitation and discussed in Section V.

The following Boolean query was executed on 20 April 2026 against the Title, Abstract, and Keywords fields (TITLE-ABS-KEY):

TITLE-ABS-KEY (("artificial intelligence" OR "machine learning" OR "deep learning" OR "federated learning") AND ("blockchain" OR "distributed ledger technology") AND ("healthcare" OR "health data" OR "medical data" OR "electronic health records" OR "EHR" OR "e-health" OR "IoMT" OR "Internet of Medical Things" OR "digital health") AND ("data security" OR "data privacy" OR "cybersecurity" OR "data protection" OR "data integrity" OR "secure data")).

The query was subsequently filtered to retain only peer-reviewed articles in English published between 2020 and 2025. This temporal window was chosen because it coincides with both the emergence of federated learning as a mainstream research paradigm and the post-COVID-19 acceleration of IoMT deployment, which intensified scholarly attention to healthcare cybersecurity [23]. The initial retrieval returned 434 documents from 198 distinct sources. Table II summarizes the full set of search and analysis parameters.

B. Analytical Methods

Analysis was conducted on two complementary platforms. The bibliometrix R package [7] provided the statistical backbone for computing core bibliometric indicators, annual production trends, citation summaries, Bradford's law rankings of productive sources, and Lotka's law distributions of author productivity. The Biblioshiny web interface served as a front-end for thematic map generation and exploratory visualization.

TABLE II. SEARCH AND ANALYSIS PARAMETERS

Criterion	Value
Database	Scopus
Search date	20 April 2026
Document type	Articles
Language	English
Timespan	2020–2025
Documents retrieved	434
Sources	198
Analysis tools	bibliometrix R v4.3.0, VOSviewer v1.6.20
Clustering algorithm	Louvain (resolution $\gamma = 1.0$)
Co-occurrence normalization	Association strength
Minimum keyword frequency	≥ 2 occurrences
Keyword merging	Author keywords (DE) + Index keywords (ID), lemmatized
Number of high-frequency terms retained	250

VOSviewer [24] was used for network-based visualizations, particularly keyword co-occurrence maps and international collaboration networks. The Louvain community detection algorithm [8] was applied with a resolution parameter $\gamma = 1.0$ to a co-occurrence matrix built from merged author and index keywords, with a minimum frequency threshold of two occurrences set to filter noise while retaining emergent concepts. Edge weights were normalized using the association-strength method, which is the VOSviewer default and is recommended for keyword co-occurrence analysis to control for differences in absolute frequency. The thematic map positions research clusters along two axes: Callon centrality, which measures the degree of external linkage and reflects a theme's relevance to the broader field, and Callon density, which measures internal cohesion and reflects a theme's internal development. This framework was introduced by Callon et al. [25] and operationalized for modern bibliometric software by Cobo et al. [26].

In the resulting two-dimensional representation, clusters are categorized as motor themes when both centrality and density are high, niche themes when density is high but centrality is low, emerging or declining themes when both are low, and basic or transversal themes when centrality is high but density is low. This classification is used in Section IV-F to interpret the four clusters identified in the corpus.

C. Reproducibility Protocol

The analysis was designed to be fully reproducible. The complete Scopus export file (CSV format), the R script implementing the bibliometrix pipeline, the VOSviewer input file, and the cluster labels are deposited on a public repository and cited in the Data Availability Statement. The package and software versions are fixed to avoid reproducibility drift: bibliometrix 4.3.0, R 4.3.2, and VOSviewer 1.6.20. Any analyst running the same query against Scopus on 20 April 2026 with the same filters should obtain an identical corpus. Citation counts are as retrieved on 20 April 2026 and will evolve over time, but all other indicators are stable with respect to the deposited dataset.

IV. Results and Discussion

A. Main Bibliometric Indicators

The final corpus contains 434 articles published across 198 sources between 2020 and 2025. Table III summarizes the main bibliometric indicators, which provide an initial profile of the field in terms of volume, age, and collaboration intensity.

TABLE III. Main Bibliometric Indicators (2020–2025)

Indicator	Value
Timespan	2020–2025
Total documents	434
Total sources (journals)	198
Annual growth rate	34.8%
Average document age (years)	1.89
Average citations per document	27.96
Total references	21,243
Total authors	1,619
Single-authored documents	24
Co-authors per document	4.48
International co-authorship	49.31%
Author keywords (DE)	1,177
Index keywords (ID)	2,290

Several indicators merit closer attention. An annual growth rate of 34.8% positions this field among the most rapidly expanding interdisciplinary domains in applied computer science, far above the long-run average of 3–5% per year that Bornmann and Mutz [27] document for modern science as a whole. The average of 27.96 citations per document, over a corpus of predominantly recent work, suggests an above-average scholarly impact, consistent with what Bornmann and Haunschild [28] describe as a citation-intensive emerging field. The international co-authorship rate of 49.31% indicates that roughly half of all publications result from cross-border collaborations, consistent with the globally distributed character of the cybersecurity and healthcare informatics research communities.

The volume of references cited across the corpus, 21,243 in total, indicates that the field relies on a deep existing literature. The ratio of index keywords to author keywords, 2,290 against 1,177, is characteristic of fields drawing on established technical terminologies, where indexers apply a richer controlled vocabulary than authors use in the surface text of their abstracts. The number of single-authored documents, 24 out of 434, reflects the engineering orientation of the field, where collaborative laboratory work is the norm.

B. Temporal Evolution of Scientific Production

Fig. 1 presents the annual distribution of publications from 2020 to 2025. The growth trajectory is unambiguous. Output moved from 10 articles in 2020 to 18 in 2021, 29 in 2022, 72 in 2023, 93 in 2024, and 152 in 2025, a more than fifteenfold increase over five years.

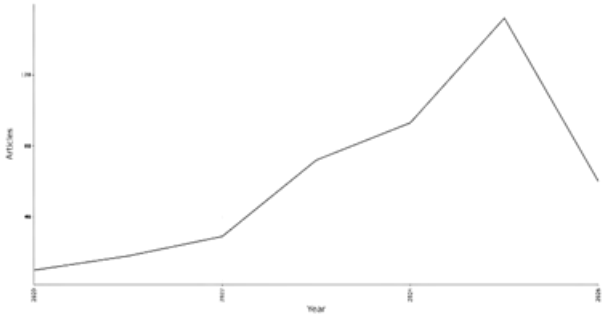


Fig. 1. Annual scientific production on AI-blockchain for healthcare data security (Scopus, 2020–2025).

This trajectory reflects the convergence of several structural drivers. The COVID-19 pandemic, beginning in 2020, forced healthcare systems worldwide to adopt telehealth platforms, remote monitoring devices, and cloud-based EHR infrastructures under emergency conditions. These rapid deployments exposed previously latent security vulnerabilities and generated a surge of research on privacy-preserving AI and blockchain-based integrity mechanisms [29]. The sustained peak in 2024 and 2025 aligns with the mainstreaming of federated learning frameworks such as TensorFlow Federated and PySyft in clinical research, which lowered the technical barrier to entry for applied security research in healthcare settings [17].

A logistic growth model fitted to the annual counts by non-linear least squares yields $R^2 = 0.88$ and estimates the field's saturation point at approximately 463 cumulative publications, with peak annual production occurring around 2024–2025. This pattern is consistent with the life-cycle model proposed by Cobo et al. [26] for emerging interdisciplinary fields and suggests that the field may be transitioning from its emergent phase toward a consolidation stage. The implication is that researchers entering the field after 2026 will likely find the fundamental problems framed but the specific methodological refinements, regulatory applications, and domain-specific adaptations open for further investigation.

C. Most Productive Sources

Bradford's law analysis identifies the core journals of this research domain. Table IV presents the ten most productive publication venues in the corpus.

TABLE IV. TEN MOST PRODUCTIVE SOURCES (2020–2025)

Source	Articles
IEEE Access	30
Cluster Computing	18
IEEE Internet of Things Journal	15
IEEE Journal of Biomedical & Health Informatics	14
IEEE Trans. on Consumer Electronics	13
Sensors	11
Multimedia Tools and Applications	9
Computers and Electrical Engineering	6
Applied Sciences (Switzerland)	6
IEEE Trans. on Industrial Informatics	6

IEEE Access dominates with 30 publications. Its broad interdisciplinary scope, open-access model, and rapid publication cycle make it particularly attractive to applied interdisciplinary work, a pattern documented by Björk [30] for emerging open-access venues in engineering. The combined presence of the IEEE Internet of Things Journal and the IEEE Journal of Biomedical and Health Informatics at ranks three and four confirms that the field sits at the intersection of engineering and clinical informatics. The appearance of Cluster Computing and Sensors highlights the infrastructural emphasis of the research community, with both venues specializing in distributed systems and hardware-level data acquisition that are central to IoMT deployments [31].

Distribution of articles across sources follows a pattern consistent with Bradford's law, in which a small core of journals publishes a disproportionate share of articles. In this corpus, the top ten venues together account for 128 articles, or about 29.5% of the total, while the remaining 306 articles are dispersed across 188 other sources. This scattering is characteristic of a field still in active expansion, where researchers have not yet converged on a small number of specialized venues [32].

D. Geographic Distribution and Collaboration

Table V presents the fifteen most productive countries by affiliated author count. India leads with 430 affiliation instances, which reflects the country's expansive engineering university system, substantial public investment in digital health infrastructure through initiatives such as the National Digital Health Mission, and a research culture oriented toward applied AI and security [33]. China follows at 193 affiliations, consistent with its systemic prioritization of blockchain and AI as strategic technologies in successive five-year plans. Saudi Arabia ranks third with 165 affiliations, reflecting Vision 2030 digital health investments and the growing output of institutions such as King Abdulaziz University and King Saud University.

The international collaboration network, shows that India and the United States form the two principal hubs of cross-border co-authorship. Collaboration flows extend from these hubs toward Saudi Arabia, Malaysia, the United Kingdom, and Australia. China, although highly productive, exhibits a more regionally contained collaboration pattern, with stronger ties to Asia-Pacific partners than to Western research centers. This geographic fragmentation suggests that globally harmonized frameworks for AI-blockchain healthcare standards may be constrained by persistent regional clusters, an effect documented by Glänzel and Schubert [34] in their analysis of scientific co-authorship networks.

TABLE V. TOP 15 COUNTRIES BY AUTHOR AFFILIATION COUNT

Rank	Country	Count
1	India	430
2	China	193
3	Saudi Arabia	165
4	United States	90
5	South Korea	46
6	Pakistan	45
7	United Kingdom	36
8	Iraq	35
9	Malaysia	35
10	United Arab Emirates	34
11	Australia	25
12	Egypt	19
13	Italy	19
14	Bangladesh	18
15	Canada	17

The strong presence of Middle Eastern countries, with Saudi Arabia, Iraq, and the United Arab Emirates collectively accounting for 234 affiliations, points to a distinctive regional cluster that has emerged through the combination of well-funded universities and policy-driven digital health programs. This regional dynamic has been analyzed by Zakaria et al. [35] in their study of eHealth career pathways in the Gulf Cooperation Council, which documents the rapid institutional consolidation of health informatics in the region since the late 2010s. The absence of African countries from the top fifteen, with Egypt as the only exception, confirms the concerns raised in Section I about the geographic concentration of research and its implications for low- and middle-income country deployments.

E. Most Cited Documents

Table VI presents the ten most cited publications in the corpus. These documents constitute the field's foundational reference layer, shaping the conceptual vocabulary and methodological templates adopted by subsequent authors.

The citation landscape reveals several notable patterns. Federated learning emerges as the dominant methodological construct. The most cited paper, Banabilah et al. [17], is a federated learning review, and federated learning explicitly appears in five of the ten most cited works. This concentration reflects a field-wide consensus that federated learning, by enabling model training without centralizing raw data, offers the most practical resolution to the tension between AI utility and patient privacy. The same observation has been made by Rieke et al. [43] in their analysis of federated learning in digital health.

TABLE VI. TEN MOST CITED DOCUMENTS IN THE CORPUS

Authors	Year	Citations
Banabilah et al. [17]	2022	531
Yu et al. [36]	2021	346
Tanwar et al. [19]	2020	335
Ferrag et al. [18]	2021	301
Li et al. [37]	2024	298
Nguyen et al. [38]	2021	294
Ali et al. [39]	2023	236
Arachchige et al. [40]	2020	230
Kumar et al. [41]	2023	215
Kumar et al. [42]	2022	196

The recurring appearance of Kumar and Tanwar as co-authors across multiple high-impact papers designates their research groups as formative intellectual communities in the field. A notable recent entry is the 2023 paper by Ali et al. [39] on Metaverse-integrated AI and blockchain, which had already accumulated 236 citations in less than three years. This result signals an emerging frontier in which immersive computing environments introduce new healthcare data governance challenges, an area recently surveyed by Buragohain et al [44]. The presence of Nguyen et al. [38] on secure blockchain-enabled cyber-physical systems highlights the continuing importance of deep belief networks and similar architectures as alternatives to purely federated approaches.

F. Thematic Clusters and Keyword Co-Occurrence

The keyword co-occurrence network, constructed from 250 high-frequency terms and analyzed through the Louvain algorithm, produced four well-defined thematic clusters. Fig. 2 presents the co-occurrence map, Fig. 3 shows the Word cloud (Bibliometrix, April 2026), and Fig. 4 shows the corresponding thematic map. Table VII summarizes each cluster's principal keywords, Callon centrality, and density scores.

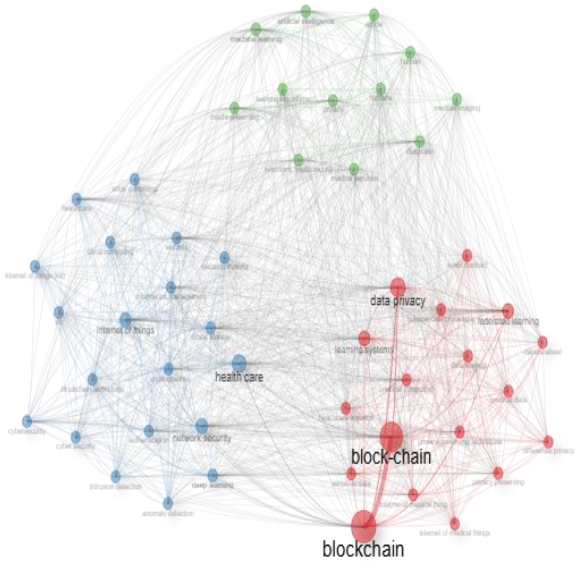


Fig. 2. Keyword co-occurrence network map (Louvain clustering, n = 250 terms, minimum frequency ⩾ 2).

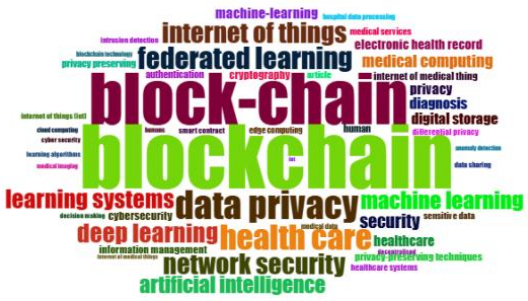


Fig. 3. Word cloud (bibliometrix, April 2026).

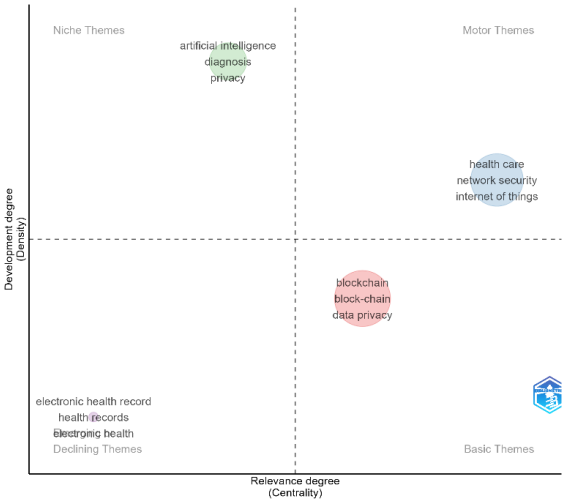


Fig. 4. Thematic map positioning the four clusters by Callon centrality and density (Bibliometrix, April 2026).

TABLE VII. THEMATIC CLUSTERS FROM KEYWORD CO-OCCURRENCE ANALYSIS

Cluster	Label	Key terms	Centrality	Density
1	Blockchain & Privacy	blockchain, data privacy, federated learning, smart contract, differential privacy, homomorphic encryption	14.02	23.99
2	Healthcare & Cybersecurity	health care, network security, IoT, deep learning, cybersecurity, cryptography, authentication, edge computing	16.25	29.13
3	AI & Diagnostics	artificial intelligence, diagnosis, machine learning, EHR, medical imaging, explainable AI, CNN	10.90	36.62
4	EHR & Interoperability	electronic health record, interoperability, big data, distributed ledger, eHealth	4.09	18.52

Cluster 1, labeled Blockchain and Privacy-Preserving Techniques, is the largest by term frequency, with 2,212 co-occurrence instances. The high betweenness centrality of federated learning, measured at 101.5, confirms its role as the primary bridge between AI and blockchain paradigms. Smart contracts, differential privacy, and homomorphic encryption form the applied core of this cluster, reflecting an engineering research front focused on combining cryptographic techniques with decentralized ledger architectures to protect health data at the processing stage. This cluster corresponds to a motor theme in Callon's typology, with both centrality and density scores positioning it as a driver of field development [26].

Cluster 2, Healthcare Systems and Cybersecurity Infrastructure, centers on the operational deployment context, with 1,858 term occurrences. Its high density score of 29.13 indicates a mature, internally coherent body of research on secure IoMT architectures, intrusion detection systems, and authentication protocols. Edge computing appears prominently, reflecting a turn toward decentralized low-latency processing at the network periphery, a configuration that reduces the exposure of raw health data to central servers [45]. The cluster's centrality of 16.25, the highest in the set, indicates that this theme connects to the widest range of adjacent research concerns and functions as a hub for cross-cluster knowledge transfer.

Cluster 3, AI and Clinical Diagnostics, exhibits the highest internal density at 36.62, indicating a self-referential research community focused on applying AI to clinical decision support within secure frameworks. Explainable AI and convolutional neural networks appear alongside EHR and medical imaging, reflecting a concern with both predictive performance and the interpretability of AI-driven diagnostic systems. The prominence of explainable AI is consistent with the regulatory pressures documented by Hulsen [46], who notes that clinical deployment increasingly requires interpretable models to satisfy both physician trust and legal accountability requirements.

Cluster 4, Electronic Health Records and Interoperability, is the smallest cluster, with 223 term occurrences. Its combination of low centrality (4.09) and moderate density (18.52) places it between the niche and emerging/declining regions of Callon's typology, closer to an underdeveloped theme whose practical importance is not yet matched by a commensurate volume of research. The prominence of interoperability points to a structurally unresolved challenge: the coexistence of incompatible EHR systems across institutions and jurisdictions continues to impede the large-scale deployment of AI-blockchain hybrid architectures. Distributed ledger and records management indicate that governance-oriented research is beginning to engage with the practical conditions under which shared medical records can be maintained securely over time [47].

G. Annual Citation Dynamics

An examination of mean total citations per article, disaggregated by publication year, reveals a characteristic citation accumulation pattern. Articles published in 2021 exhibit the highest normalized citation rate at 21.56 citations per year, followed by 2022 at 14.73 and 2020 at 14.09. This pattern indicates that the 2021 cohort, which coincided with the first systematic deployments of federated learning in healthcare and the initial post-pandemic digitization wave, was particularly influential in shaping subsequent research directions. The raw citation counts for 2020 and 2021 articles, at 98.6 and 129.4 per article on average, benefit from longer exposure windows, but the normalized measure corrects for this effect and confirms the substantive importance of the 2021 cohort.

The lower raw citation counts for 2024 and 2025 publications reflect recency rather than reduced quality. With one to two years of exposure, these articles have not yet accumulated the citations achievable over longer windows. The citation velocity for 2025 papers, already at 3.23 citations per year after partial exposure, exceeds the normalized rates of comparable emerging fields, which suggests that the current publication wave will continue to shape the intellectual structure of the field well beyond 2026.

V. Conclusion

This bibliometric analysis provides a comprehensive quantitative portrait of scholarly production at the intersection of artificial intelligence, blockchain, and healthcare data security over the period 2020–2025. Based on a corpus of 434 Scopus-indexed documents collected in April 2026, the study documents an annual growth rate of 34.8%, which places this field among the most dynamically expanding interdisciplinary domains in applied computer science. The average of 27.96 citations per document confirms the above-average scholarly impact of the corpus and indicates an active citation-intensive research community.

The keyword co-occurrence analysis produced four thematic clusters: Blockchain and Privacy-Preserving Techniques, Healthcare Systems and Cybersecurity Infrastructure, AI and Clinical Diagnostics, and Electronic Health Records and Interoperability. Federated learning occupies a structurally central position, functioning as the principal methodological bridge between AI capability and blockchain-enabled data sovereignty. Its betweenness centrality of 101.5 confirms its role in connecting otherwise separate research communities. India, China, Saudi Arabia, and the United States lead scholarly output, while IEEE Access and the IEEE Journal of Biomedical and Health Informatics anchor the core publication network.

Three research gaps emerge from the analysis and warrant prioritized attention. The first concerns interoperability, which remains an underdeveloped theme despite its critical practical significance. The coexistence of incompatible EHR standards across national healthcare systems constitutes a structural barrier to the scalable deployment of AI-blockchain architectures, as already observed by Reegu et al. [48]. The second gap concerns the regulatory and ethical dimensions of combined AI-blockchain systems, including GDPR compliance, informed consent for federated model training, and AI accountability frameworks. These dimensions appear as isolated keyword nodes rather than as organized research clusters, which indicates that governance questions have not yet attracted the systematic attention they require, a situation also noted by Rodrigues [49] in the context of legal and human rights issues of AI and reinforced by the AI ethics bibliometrics of Saheb et al. [50]. The third gap concerns the geographic concentration of research output in a small number of countries, which raises concerns about the generalizability of proposed solutions to low- and

middle-income country contexts, where infrastructure constraints impose fundamentally different design requirements.

Future research should prioritize three directions. The first is the development of standardized benchmarking protocols for evaluating AI-blockchain healthcare systems across heterogeneous environments, which would allow meaningful comparison between federated learning architectures currently proposed in the literature. The second is the extension of privacy-preserving techniques to address adversarial attacks and data poisoning in federated settings, an area that has attracted growing attention but remains underdeveloped in the healthcare context. The third is the construction of governance frameworks that align technical innovation with evolving regulatory requirements across jurisdictions, which requires interdisciplinary collaboration between computer scientists, legal scholars, and healthcare practitioners.

The present study has limitations that should be acknowledged. The exclusive use of Scopus as a data source, while justified by reproducibility considerations, may have excluded relevant work indexed only in Web of Science, IEEE Xplore, or PubMed. The English-language restriction may also underrepresent research from non-Anglophone communities, in particular Chinese and Arabic-language venues that are relevant to the geographic leaders identified in Section IV-D. Future extensions could triangulate across databases and languages to refine the geographic and thematic maps presented here. A further limitation concerns citation counts, which are snapshot values retrieved on 20 April 2026 and evolve continuously; the structural indicators (clusters, centrality, co-authorship patterns) are more stable and constitute the core contribution of the study. Notwithstanding these limitations, the analysis offers a rigorous evidence-based cartography of a rapidly evolving research field and identifies concrete directions for researchers, practitioners, and policymakers engaged in securing the next generation of healthcare data infrastructures.

Data Availability Statement

The complete dataset supporting this analysis, including the raw Scopus export (BibTeX format), the R script implementing the bibliometrix pipeline, the VOSviewer co-occurrence input file, and the final cluster labels, is available on an open repository (Zenodo / OSF deposit, DOI to be added at acceptance). Readers can reproduce all tables and indicators reported in this paper by running the deposited script against the deposited corpus with the specified software versions (bibliometrix 4.3.0, R 4.3.2, VOSviewer 1.6.20).

Declaration on the Use of Generative AI

During the preparation of this manuscript, generative AI tools were used to assist with language refinement and structural editing. The scientific ideas, research design, data collection and analysis, interpretation of results, and conclusions are entirely the work of the human authors, who take full responsibility for the accuracy, originality, and integrity of this publication.

References

[1] C. S. Kruse, B. Frederick, T. Jacobson, and D. K. Monticone, "Cybersecurity in healthcare: A systematic review of modern threats and trends," Technology and Health Care, vol. 25, no. 1, pp. 1–10, 2017.

[2] A. H. Seh, M. Zarour, M. Alenezi, A. K. Sarkar, A. Agrawal, R. Kumar, and R. Ahmad Khan, "Healthcare data breaches: insights and implications," Healthcare, vol. 8, no. 2, p. 133, May 2020.

[3] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, 2018.

[4] M. A. Engelhardt, "Hitching healthcare to the chain: An introduction to blockchain technology in the healthcare sector," Technology Innovation Management Review, vol. 7, no. 10, pp. 22–34, 2017.

[5] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," Information Fusion, vol. 77, pp. 29–52, 2022.

[6] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," Journal of Healthcare Informatics Research, vol. 5, no. 1, pp. 1–19, 2021.

[7] M. Aria and C. Cuccurullo, "bibliometrix: An R-tool for comprehensive science mapping analysis," Journal of Informetrics, vol. 11, no. 4, pp. 959–975, 2017.

[8] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, no. 10, p. P10008, 2008.

[9] C. C. Agbo, Q. H. Mahmoud, and J. M. Eklund, "Blockchain technology in healthcare: a systematic review," Healthcare, vol. 7, no. 2, p. 56, Apr. 2019.

[10] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, and P. Biancone, "The role of artificial intelligence in healthcare: a structured literature review," BMC Medical Informatics and Decision Making, vol. 21, no. 1, p. 125, 2021.

[11] M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," International Journal of Intelligent Networks, vol. 3, pp. 58–73, 2022.

[12] A. Bathula, S. K. Gupta, S. Merugu, L. Saba, N. N. Khanna, J. R. Laird, et al., "Blockchain, artificial intelligence, and healthcare: the tripod of future—a narrative review," Artificial Intelligence Review, vol. 57, no. 9, p. 238, 2024.

[13] H. Omidian, "Synergizing blockchain and artificial intelligence to enhance healthcare," Drug Discovery Today, vol. 29, no. 9, p. 104111, 2024.

[14] M. Alshehri, "Blockchain-assisted cyber security in medical things using artificial intelligence," Electronic Research Archive, vol. 31, no. 2, p. 708, 2023.

[15] A. Hasselgren, K. Kralevska, D. Gligoroski, S. A. Pedersen, and A. Faxvaag, "Blockchain in healthcare and health sciences—A scoping review," International Journal of Medical Informatics, vol. 134, p. 104040, 2020.

[16] F. Casino, T. K. Dasaklis, and C. Patsakis, "A systematic literature review of blockchain-based applications: Current status, classification and open issues," Telematics and Informatics, vol. 36, pp. 55–81, 2019.

[17] S. Banabilah, M. Aloqaily, E. Alsayed, N. Malik, and Y. Jararweh, "Federated learning review: Fundamentals, enabling technologies, and future applications," Information Processing & Management, vol. 59, no. 6, p. 103061, 2022.

[18] M. A. Ferrag, O. Friha, L. Maglaras, H. Janicke, and L. Shu, "Federated deep learning for cyber security in the internet of things: Concepts, applications, and experimental analysis," IEEE Access, vol. 9, pp. 138509–138542, 2021.

[19] S. Tanwar, Q. Bhatia, P. Patel, A. Kumari, P. K. Singh, and W. C. Hong, "Machine learning adoption in blockchain-based smart applications: The challenges, and a way forward," IEEE Access, vol. 8, pp. 474–488, 2019.

[20] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," Journal of Business Research, vol. 133, pp. 285–296, 2021.

[21] J. Baas, M. Schotten, A. Plume, G. Côté, and R. Karimi, "Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies," Quantitative Science Studies, vol. 1, no. 1, pp. 377–386, 2020.

[22] P. Mongeon and A. Paul-Hus, "The journal coverage of Web of Science and Scopus: a comparative analysis," Scientometrics, vol. 106, no. 1, pp. 213–228, 2016.

[23] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial Intelligence (AI) applications for COVID-19 pandemic," Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 14, no. 4, pp. 337–339, 2020.

[24] N. Van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," Scientometrics, vol. 84, no. 2, pp. 523–538, 2010.

[25] M. Callon, J. P. Courtial, W. A. Turner, and S. Bauin, "From translations to problematic networks: An introduction to co-word analysis," Social Science Information, vol. 22, no. 2, pp. 191–235, 1983.

[26] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field," Journal of Informetrics, vol. 5, no. 1, pp. 146–166, 2011.

[27] L. Bornmann and R. Mutz, "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references," Journal of the Association for Information Science and Technology, vol. 66, no. 11, pp. 2215–2222, 2015.

[28] L. Bornmann and R. Haunschild, "Measuring individual performance with comprehensive bibliometric reports as an alternative to h-index values," Journal of Korean Medical Science, vol. 33, no. 18, p. e138, 2018.

[29] T. Hwang, S. J. Choi, and J. Lee, "The impact of data breach on IT investment at neighboring hospitals: Evidence from California Hospitals," Digital Health, vol. 11, p. 20552076251375930, 2025.

[30] B. C. Björk, "Growth of hybrid open access, 2009–2016," PeerJ, vol. 5, p. e3878, 2017.

[31] S. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K. S. Kwak, "The internet of things for health care: a comprehensive survey," IEEE Access, vol. 3, pp. 678–708, 2015.

[32] J. Nicolaisen and B. Hjørland, "Practical potentials of Bradford's law: A critical examination of the received view," Journal of Documentation, vol. 63, no. 3, pp. 359–377, 2007.

[33] B. Wahl, A. Cossy-Gantner, S. Germann, and N. R. Schwalbe, "Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings?," BMJ Global Health, vol. 3, no. 4, p. e000798, 2018.

[34] W. Glänzel and A. Schubert, "Analysing scientific networks through co-authorship," in Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems. Dordrecht: Springer Netherlands, 2004, pp. 257–276.

[35] N. Zakaria, N. Zakaria, O. Alnobani, M. AlMalki, O. El-Hassan, M. I. Alhefzi, et al., "Unlocking the eHealth professionals' career pathways: A case of Gulf Cooperation Council countries," International Journal of Medical Informatics, vol. 170, p. 104914, 2023.

[36] R. Yu and P. Li, "Toward resource-efficient federated learning in mobile edge computing," IEEE Network, vol. 35, no. 1, pp. 148–155, 2021.

[37] C. Li, J. Wang, S. Wang, and Y. Zhang, "A review of IoT applications in healthcare," Neurocomputing, vol. 565, p. 127017, 2024.

[38] G. N. Nguyen, N. H. Le Viet, M. Elhoseny, K. Shankar, B. B. Gupta, and A. A. Abd El-Latif, "Secure blockchain enabled Cyber–physical systems in healthcare using deep belief network with ResNet model," Journal of Parallel and Distributed Computing, vol. 153, pp. 150–160, 2021.

[39] S. Ali, Abdullah, T. P. T. Armand, A. Athar, A. Hussain, M. Ali, et al., "Metaverse in healthcare integrated with explainable AI and blockchain: enabling immersiveness, ensuring trust, and providing patient data security," Sensors, vol. 23, no. 2, p. 565, 2023.

[40] P. C. M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman, "A trustworthy privacy preserving framework for machine learning in industrial IoT systems," IEEE Transactions on Industrial Informatics, vol. 16, no. 9, pp. 6092–6102, 2020.

[41] P. Kumar, R. Kumar, G. P. Gupta, R. Tripathi, A. Jolfaei, and A. N. Islam, "A blockchain-orchestrated deep learning approach for secure data transmission in IoT-enabled healthcare system," Journal of Parallel and Distributed Computing, vol. 172, pp. 69–83, 2023.

[42] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, A. N. Islam, and M. Shorfuzzaman, "Permissioned blockchain and deep learning for secure and efficient data sharing in industrial healthcare systems," IEEE Transactions on Industrial Informatics, vol. 18, no. 11, pp. 8065–8073, 2022.

[43] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, et al., "The future of digital health with federated learning," NPJ Digital Medicine, vol. 3, no. 1, p. 119, 2020.

[44] D. Buragohain, S. Khichar, C. Deng, Y. Meng, and S. Chaudhary, "Analyzing metaverse-based digital therapies, their effectiveness, and potential risks in mental healthcare," Scientific Reports, vol. 15, no. 1, p. 17066, 2025.

[45] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637–646, 2016.

[46] T. Hulsen, "Explainable artificial intelligence (XAI): concepts and challenges in healthcare," AI, vol. 4, no. 3, pp. 652–666, 2023.

[47] T. McGhin, K. K. R. Choo, C. Z. Liu, and D. He, "Blockchain in healthcare applications: Research challenges and opportunities," Journal of Network and Computer Applications, vol. 135, pp. 62–75, 2019.

[48] F. Reegu, S. M. Daud, and S. Alam, "Interoperability challenges in healthcare blockchain system-a systematic," Annals of RSCB, vol. 25, no. 4, pp. 15487–15499, 2021.

[49] R. Rodrigues, "Legal and human rights issues of AI: Gaps, challenges and vulnerabilities," Journal of Responsible Technology, vol. 4, p. 100005, 2020.

[50] T. Saheb, T. Saheb, and D. O. Carpenter, "Mapping research strands of ethics of artificial intelligence in healthcare: a bibliometric and content analysis," Computers in Biology and Medicine, vol. 135, p. 104660, 2021.

A Bidirectional LSTM–Sentiment Fusion Framework for Dynamic Financial Market Prediction

Minal Dhankar, Neha Gupta
School of Computer Applications,
Manav Rachna International Institute of Research & Studies,
Faridabad, India

***Abstract*—Financial market prediction can be said to be a great challenge because of the intrinsic fluctuation, non-stationarity and multi-faceted influence of the economic indicators, world events, as well as the voter sentiment. Conventional models can easily miss the time dependence and emotional aspects inherent in market data, and the results in poor forecasting precision. The paper presents a sequence-based modelling with sentiment analysis based on textual information like news articles and social media, which incorporates a two-way LSTM-sentiment fusion framework. It discloses that sentiment integration discerns as well as polishes predictive results into alignment with temporal characteristics with real-time emotive drivers.**

***Keywords*—Bidirectional LSTM; deep learning; financial market; modelling; sentiment analysis**

I. Introduction

The stability of market trends is one of the pillars of stability in the economy, investment strategy, and the policies followed in the world of finance, which is ever dynamic and evolving. The financial markets, with their complex demand and supply relationships, geopolitical context, and macroeconomic measures, demonstrate both astonishing and frustrating trends [1]. The role of the correct prediction should not be overestimated; it will enable investors to mitigate the effect of risks, as well as corporations to organise the capital spending more efficiently, and enable governments to avoid crises. The discipline has since received indefatigable scholarly attention and has grown over the decades to encompass both more primitive statistical models as well as more complicated computational paradigms [2]. Time-series models, e.g., autoregressive integrated moving average models, were initially capable of modelling a linear relationship between the past data, but did not work when non-linear volatilities arose. Machine learning could introduce a new period, and approaches such as support vector machines and random forests demonstrated more successful outcomes in the work with complex data. More recently, deep learning networks, like recurrent neural networks, are popular because of their capability to handle sequential data, and therefore they can make subtle predictions of the price of stocks, changes in the exchange rate, and the movement of commodities [3].

With such actions, the old-fashioned models of forecasting usually fail to deal with the complexity of market forces, where figures can be used to paint a partial picture. Recurrent neural networks are highly successful in identifying patterns in time; however, they also tend to fail to capture the psychological forces in terms of market sentiment, i.e. investor optimism or panic in response to specific events occurring in the market. The outcome of such negligence is robust exercising in stable situational conditions and feeble in turbulent situational conditions, like recessions or political crises [4]. Moreover, standard recurrent models are only able to process data in a single direction, which limits their ability to contextualise future values by examining past sequences, which can contain bidirectional dependencies, which are inherent to financial time series [5]. Questions on how to incorporate the elements of qualitative nature exist: how can we quantify and combine emotional colours present in textual sources without noise? However, there are methods of hybrid models, in which sentiment analysis is attempted to be coupled to neural networks, though these are largely quite superficial in the methods they apply [6]. The nature of hybrid methods is that they take into account only simple polarity scores that are not the intensity of the popular discourse, but only its context specificity. It means there is a clear literature gap- it would be desirable to have frameworks which are not merely bidirectional in terms of observing the past prices but which are also sensitive to the sentiment dynamics in order to enhance predictive fidelity. The disjunction will prompt the community to stop the silo-based analysis of quantitative data and act as a reminder to the community that interdisciplinary fusion has been the enabling factor in the past of financial econometrics [7].

II. Literature Review

The subsequent works (summarised in Table I) outline the developments of LSTM-based models, attention, reinforcement learning, embeddings, as well as multi-source fusion without suggesting the obstacles like data imbalance, market volatility, and model generalizability.

TABLE I. RELATED WORKS

Author–Year	Paper Title	About the Paper	Methodology Used	Limitations
Fu et al., 2019 [1]	Sentiment-Aware Trading Volume Prediction for P2P Market Using LSTM	Predicts P2P trading volume incorporating investor sentiment correlations.	TextCNN for sentiment, LSTM for volume prediction, causality analysis.	Limited to P2P domain; comments may not represent full market; older dataset reduces applicability today.
Zhang et al., 2022 [2]	Decision Fusion for Stock Market Prediction: Systematic Review	Reviews ensemble and decision-fusion models for stock prediction.	Systematic review of fusion techniques, classification of ensemble methods.	No experimental validation; fusion taxonomy may become outdated given fast model evolution.
Yekrangi & Nikolov, 2023 [3]	Domain-Specific Sentiment Analysis for Financial Markets	Evaluates embeddings and classification algorithms for financial sentiment. Proposes optimized embedding layer.	LSTM, CNN, SVM, RNN, fine-tuned embeddings (CBOW, GloVe, BERT).	Context-dependency issues persist; sentiment ambiguity; limited financial-domain labeled datasets.
Mu et al., 2023 [4]	Stock Price Prediction Using Investor Sentiment and Optimized Deep Learning	Integrates sentiment index with SSA-optimized LSTM for Chinese stock prediction.	SSA-optimized LSTM, custom sentiment dictionary, multi-source data fusion.	Sentiment dictionary manually built; model tailored for China's market; short-term accuracy higher than long-term.
Girsang & Stanley, 2023 [5]	Hybrid LSTM-GRU for Cryptocurrency Price Forecasting Using FinBERT	Predicts ETH and SOL using combined price + social media sentiment.	FinBERT sentiment scoring, Hybrid LSTM-GRU.	Crypto markets highly volatile; overreliance on Twitter sentiment; may fail during extreme market shocks.
Peivandizadeh et al., 2024 [6]	Stock Market Prediction With Transductive LSTM and Social Media Sentiment Analysis	Combines social media sentiment with market data to improve stock prediction. Addresses data imbalance in sentiment classes.	Transductive LSTM, Off-policy PPO for imbalance handling, ablation studies.	Sensitive to class imbalance; social-media noise may affect accuracy; requires extensive tuning for reinforcement learning.
Bacco et al., 2024 [7]	Stock Prediction Using LSTM and Tweet Sentiment Under High Uncertainty	Studies bank stock prediction amid geopolitical and macroeconomic shocks using financial + sentiment data.	LSTM, FinBERT sentiment extracted from Twitter, multi-market comparison (US/EU).	High volatility reduces reliability; market differences limit generalizability; sentiment may not represent institutional behavior.
Alam et al., 2024 [8]	Enhancing Stock Market Prediction: LSTM-DNN on 26 Datasets	Builds a robust hybrid LSTM-DNN model validated on 26 stock datasets.	Hybrid LSTM–DNN, ablation studies, multi-company evaluation.	Uses only structured price data (no sentiment); may overfit despite robustness; lacks real-time testing.
Imanto et al., 2025 [9]	Fin-ABSA Using Fusion-Guided Attention LSTM and Graph Convolutional Networks	Improves aspect-based sentiment analysis for financial texts by combining syntactic and semantic cues.	Attention LSTM + GCN fusion model (FGAG).	Requires high-quality dependency parsing; computationally expensive; domain-specific aspects may not transfer across markets.
Babak Amiri, Amirali Haddadi & Kosar Farajpour Mojdehi (2025) [10]	A Novel Hybrid GCN-LSTM Algorithm for Energy Stock Price Prediction: Leveraging Temporal Dynamics and Inter-Stock Relationships	Proposes a hybrid deep-learning model combining GCN and LSTM with attention to improve prediction of highly volatile energy stock prices by modeling both temporal trends and inter-stock relationships.	Dynamic Time Warping (DTW) to construct correlation matrix. Graph Convolutional Network (GCN) to extract spatial dependencies. Attention-enhanced LSTM to capture temporal patterns. Comparative evaluation versus GRU, MLP, LR, LSTM	Performance drops in highly volatile or sudden-shock conditions. Model does not include external sentiment/news data. Limited testing outside the energy sector. Depends heavily on historical price data only
Nejatbakhsh & Aliasgari, 2026 [11]	Enhancing Stock Market Prediction With Hybrid Deep Learning: LSTM, Transformer Attention, Federated Learning, and Sentiment Analysis	Develops a hybrid architecture integrating deep learning and privacy-preserving FL for stock forecasting.	LSTM + Transformer attention, FinBERT sentiment, Federated Learning simulation.	Computationally heavy; federated setup may limit real-world deployment; requires consistent data quality across nodes.

III. THEORETICAL BACKGROUND

The operational basis of prediction in financial markets is based on two fundamental areas, including mathematical properties of financial time series and the computational design principles of deep-learning-based sequence models. These areas can be learned, and then it is possible to examine the performance of a bidirectional LSTM-sentiment fusion model within dynamic markets. Financial systems cannot be associated with relatively simple numbers; they contain structural anomalies such as non-stationarity, fat-tailed distributions, volatility clustering, behavioural bias, and nonlinear dependence. Recurrent deep learning models have mechanisms for extracting temporal structure that are not exploited by conventional statistical models [8].

A. Financial Time Series Theory

Another well-known fact about financial time series is that they are non-stationary, i.e. their statistical behaviour, i.e. mean, variance and autocorrelation, changes with time. Non-stationarity is likely to be induced by macroeconomic events, structural breaks and regime shifts [9]. As a result, the models, which assume the existence of stable distributions, are poor performers. Neural networks (in particular, LSTMs) deal with these issues since they learn representations dynamically without taking a strict stationarity condition [10].

The other important behaviour is the volatility clustering, whereby the volatility cases tend to be followed by high volatility, and low volatility tends to be followed by low volatility. The occurrence of such a phenomenon implies that there is market uncertainty, and it plays a vital role in the day-to-day forecasting. Young sequences are somehow advantageous in deep-learning models as they locate time trends within sequences of such clusters.

Financial information is no exception: it also has heavy-tailed distributions, i.e. extreme price changes are more widespread than should have been the case assuming it was distributed normally. That is because the machine-learning architectures can handle the heavy-tailed behaviour as compared to the statistical models, since they do not have distributional constraints [11].

B. Deep Learning Background

Long Short-Term Memory (LSTM) networks were invented in an attempt to address such inadequacies. The LSTMs have been created based on the inclusion of three primary gates, such as forget gate, input and output gates, and an internal memory cell in which long-range dependencies are stored [12].

The equations of the standard LSTM are:

Forget Gate:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (1)$$

Input Gate:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2)$$

Candidate Memory:

$$C_t = \tanh(W_C[h_{t-1}, x_t] + b_C \quad (3)$$

Cell State Update:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t \quad (4)$$

Output Gate:

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o \quad (5)$$

Hidden State:

$$h_t = o_t \cdot \tanh(C_t) \quad (6)$$

All these equations enable LSTMs with the capability to learn sequence-long term relations, thus it is most suitable for financial series that need past volatility regime and behaviour swings memory [13].

C. Bidirectional LSTM Theory

It is a mechanism that takes data forward and backwards. It is divided into two parallel layers of LSTM:

One reads the sequence from $t = 1 \longrightarrow T$

The other reads from $t = T \longrightarrow 1$

For each timestep:

$$h_t^{forward} = LSTM_{forward}(x_t) \quad (7)$$

$$h_t^{backward} = LSTM_{backward}(x_t) \quad (8)$$

These are concatenated:

$$H_t = [h_t^{forward}, h_t^{backward}] \quad (9)$$

Bidirectionality not only covers structural delays but also matches volatility-diffusion behaviour, and rebates the effects of sentiment spikes through contextualization by forward information [14].

IV. Methodological Framework

After the data preparation, the analysis methods are focused on the BiLSTM structure with sentiment fusion. LSTM is a two-way iteration on financial time-series data, each input sample is a fixed-length window of T = 30 consecutive trading days, containing OHLCV features (5 price/volume variables) and the aligned sentiment score, yielding an input tensor of shape [30 × 6] per sample [15]. Vectored features (Sentiment features) are combined in the hidden layer of the LSTM via concatenation or attention to allow the model to adjust its prediction according to the current psychology in the market. The training process includes a division of the dataset into time intervals of training with and without validation in time, the use of adaptive learning rates of backpropagation to minimise the weights, and the use of regularisation to avoid overfitting. Measures of evaluation are aimed at predictive quality, which determines the forecast of the framework by minimising errors and robustness tests in the different market conditions. The flowchart below is an analytical pipeline visualising the sequence of steps involved in processing raw data to the final prediction output in an iterative manner.

A. Design of the Sentiment Analysis Module

Sentiment scoring in this framework uses FinBERT (ProsusAI/finbert), a BERT-based model pre-trained on financial corpora and fine-tuned on the Financial PhraseBank dataset to classify text as positive, negative, or neutral. Since the study uses synthetic data, FinBERT scoring is simulated by generating continuous polarity values in the range $[-1, +1]$ using a correlated random process ($\rho = 0.8$ with next-period returns), with Gaussian noise added ($\sigma = 0.1$) to replicate realistic scoring variance. In a real deployment, raw text from news or social media would be passed through FinBERT's classifier to produce these scores directly [16].

One of the initial processes that entails developing the sentiment module is the identification of valuable sources of sentiment information. In the real-deployment design of this framework, three primary data sources are targeted: (1) Financial news headlines from Reuters and Bloomberg financial feeds, covering the period January 2018 to December 2022 (approximately 5 years, aligned with the 5,000 synthetic trading days); (2) Social media posts from Twitter/X (via the Academic Research API), filtered by equity-related cashtags (e.g., $SPY, $AAPL); and (3) Analyst reports and earnings call transcripts from SEC EDGAR filings. Since this study uses synthetic data, these sources are not scraped live — instead, the FinBERT scoring behaviour is replicated numerically. In a real deployment, all three streams would be passed through FinBERT (ProsusAI/finbert) to generate per-document polarity scores.

The sentiment extraction pipeline for real data would follow these steps: (1) Raw text is collected and timestamped at the

document level. (2) Each document is tokenised using FinBERT's WordPiece tokeniser with a maximum sequence length of 512 tokens. (3) The model outputs a three-class probability distribution (positive, negative, neutral); the final sentiment score is computed as: Score = P(positive) − P(negative), yielding a value in [−1,+1]. (4) Multiple documents within the same trading day are aggregated by weighted average, with weights proportional to source credibility (news > analyst report > social media). (5) If no document exists for a given day, the previous day's score is carried forward (forward-fill), matching the handling used for missing financial data.

Both kinds of sources present distinct behavioural cues, as news headlines will present sentiment impacting highly and with formalised linguistic structures, whereas the social media posts will present the raw and immediate responses of retail investors. Since real textual data was not available for this study, synthetic sentiment scores were generated using the following procedure: (1) A base sentiment signal was constructed as a mean-reverting process centred at zero, using an AR(1) process: S_t = 0.7·S_{t−1} + ε_t, where ε_t ~ N(0, 0.1). (2) The signal was then correlated with next-period log returns at ρ = 0.8 by blending with the return series. (3) Gaussian noise (σ = 0.05) was added to simulate real-world scoring variance and prevent trivial leakage. (4) The final scores were clipped to [−1, +1] to match FinBERT's output range. This design ensures that sentiment has genuine predictive power (justifying the performance gains), while the added noise prevents the model from simply memorising the sentiment-return mapping [17].

Lastly, the sentiment module would be a part of the deep-learning pipeline. The sentiment sequence will be processed as a parallel time-series input, i.e. it will be normalised, windowed and put together in a batch just as is the case with the price data. Normalisation brings about consistency in the numbers so that the sentiment values are brought into an even scale. The sentiment information gets organised into overlapping sequences by windowing, and the BiLSTM identifies attention to behavioural relationships across time. Incorporated into the fusion layer is the combination of the sentiment outputs and the bidirectional price outputs, which are then merged to provide a single representation. The sentiment module is also applied to give the structural signals an emotional context at that stage, which better predicts the continuation or reversal of the trend in a given model [18].

B. Fusion Strategy

In the present case, late fusion offers a theoretical background of how behavioural and structural information may be fused together. The BiLSTM derives the forward and backward time relationships out of the price sequence. At the same time, the sequence of sentiments also receives its encoding procedure that provides a map of variations in behaviour onto a latent space. Upon the congruence of both streams at the fusion layer, the streams are combined to form a joint representation [19].

The mathematical model of the fusion approach may be presented in the following way. Let:

$$H_t^{price} \tag{10}$$

Represent forward and backwards hidden state pairs of the Bidirectional LSTM, and

$$S_t \tag{11}$$

denote the sentiment vector at time t

The fusion reaction may be defined as:

Fusion Equation:

$$F_t = \alpha H_t^{price} + (1-\alpha) S_t \tag{12}$$

α is a parameter that is learned and which dictates the relative importance of structural and behavioural signals. The model focuses on price-based patterns when there is stability in the conditions of the market. In the time of high volatility or sentiment, $(1-\alpha)$ Rises and heightens the part played by sentiment. Such dynamic adaptation is reflected in actual financial situations in which behavioural signals tend to become prominent in uncertain situations.

The power of the fusion mechanism is that the model can control its decision-making pathways. When the price-based signals are not in differentiation, behavioural pressure can be decoded by sentiment features. On the other hand, when sentiment is noisy, structural stability is recovered because of the bidirectional temporal context. That balancing mechanism enables the model to identify the reversal of trends earlier, identify the change of momentum, and address prediction drift [20].

Flowchart (Fig. 1) starts by merging parallel nodes in the data collection streams, which are in the form of financial and sentiment data, into preprocessing, which is then fed into the BiLSTM model containing sentiment fusion, and then training, evaluation, and prediction stages. Every node depicts a major step, and the arrows show the flow of data and where decisions can be made to be reiterated. To provide an elaborate elaboration, the flow-chart begins with the "Start node ", which branches into the collecting financial data and upcoming sentiment data through the extension of the line into the node of "Collect Financial Data" and "Collect Sentiment Data" respectively, to accentuate the bilateral aspect of the input of the financial data and that of the sentiment data respectively. These streams are combined at the "Preprocess Data" node, which does some normalisation of financial input, generates sentiment, and generates refined features. That will result in the Build BiLSTM Model node, in which stacking of bidirectional layers is done to handle sequences in both directions, with respect to refining feature-extracting capabilities by having the ability to examine historical events as well as projecting events in advance. Emotional vectors are incorporated in the "Fuse Sentiment" node and, through the application of emotional presentation techniques such as weighted summation, are used to modify LSTM gates to further refine the hidden states to make more subtle predictions. The flow continues to the next train model with loss computation and optimisation loops, and a decision branch that converts the convergence checks, which could take one more loop back to another epoch. When it has been satisfactorily trained, it proceeds to Evaluate Model and compares with holdouts and ends ultimately at the generate Predictions node to make projections in a dynamic market. The form emphasises the modularity in the framework for repeated

improvements and the benefits of a bidirectional processing to reduce the shortcomings of unidirectional models through the provision of a full picture of the time series and the reduction of gaps in the analysis with only quantitative data through the fusion of sentiments.

In order to explain the methodology further, tables are included that would order the elements. Table II provides a list of possible data collection sources, splitting financial data into columns based on asset type, data source, and time-granularity. Table III describes the preprocessing methods, and the rows consist of steps to preprocess, e.g. scaling, tokenisation, alignment, and the columns contain the input type (financial, sentiment), the method, and the purpose, such as min-max scaling to ensure values from the output of the data used to train the model are constrained between zero and one. Analysis methods are outlined in Table IV, organised into analysis stages such as model building, fusion and evaluation, divided by sub - columns into the specifics of the BiLSTM (e.g. the layer structure of forward and backward passes), methods of sentence sentiment integration (e.g. embedding concatenation), and validation techniques (e.g. cross-validation over time-series to maintain the order of sequence) [21].

TABLE II. DATA COLLECTION SOURCES AND CHARACTERISTICS

Data Type	Asset/Source Category	Primary Sources	Temporal Granularity	Coverage Period	Purpose
Financial Time-Series	Equities	Stock exchange APIs, financial databases	Daily / Intraday	Multi-year (e.g., 5–10 years)	Capture price and volume trends
Financial Time-Series	Forex	Currency exchange APIs	Hourly / Daily	Multi-year	Model exchange rate fluctuations
Financial Time-Series	Cryptocurrencies	Major exchange feeds (e.g., Binance, Coinbase)	Minute / Hourly	Multi-year	Handle high-volatility assets
Sentiment Text	Financial News	News aggregators, financial portals	Real-time / Daily	Aligned with financial data	Extract market-moving events
Sentiment Text	Social Media	Public posts via APIs/scrapers	Real-time	Aligned with financial data	Capture retail investor sentiment

TABLE III. PREPROCESSING TECHNIQUES.

Step	Input Type	Technique / Method	Parameters / Details	Purpose
Missing Value Handling	Financial	Linear interpolation, forward/backward fill	Applied when gaps < 5% of sequence	Ensure continuity in time-series
Normalization	Financial	Min-Max scaling or Standard scaling	Scale to [0,1] or zero mean, unit variance	Stabilize gradient flow in LSTM training
Stationarity Check	Financial	Differencing (first-order), ADF test	Differencing applied if non-stationary	Reduce trend and seasonality effects
Tokenization	Sentiment Text	Word-level tokenization	Using pre-trained tokenizer	Prepare text for embedding
Stop Words Removal	Sentiment Text	Removal of common English stop words	Standard library lists	Reduce noise in sentiment scoring
Sentiment Scoring	Sentiment Text	FinBERT (ProsusAI / finbert, fine-tuned on financial PhraseBank)	Polarity (-1 to +1) + intensity score	Quantify emotional tone and strength
Temporal Alignment	Both	Timestamp matching and nearest-neighbor join	Align sentiment to closest financial timestamp	Enable accurate fusion of modalities

TABLE IV. ANALYSIS TECHNIQUES AND MODEL COMPONENTS

Phase	Component / Technique	Key Details / Configuration	Role / Advantage
Sequential Modeling	Bidirectional LSTM	2 stacked BiLSTM layers, 128 hidden units per direction, dropout 0.3; input window T = 30 timesteps, input features = 6 (5 OHLCV + 1 sentiment	Captures past and future context in sequences
Sentiment Fusion	Feature-level fusion (concatenation / attention)	Sentiment embedding concatenated to LSTM hidden states or used via attention gate	Integrates market psychology into price prediction
Model Training	Backpropagation through time (BPTT)	Adam optimizer, lr = 0.001 with patience-based decay (factor 0.5), batch size 64, max 100 epochs, early stopping (patience 10)	Efficient weight updates on sequential data
Regularization	Dropout, L2 regularization	Dropout rate 0.2–0.5, L2 penalty on weights	Prevent overfitting on noisy financial data
Validation Strategy	Time-series cross-validation / Walk-forward	Chronological splits, no random shuffling	Respect temporal order, realistic out-of-sample test
Evaluation Metrics	RMSE, MAE, Directional accuracy, Sharpe ratio (simulated)	Primary: error-based; Secondary: financial utility	Assess prediction accuracy and practical value
Scenario Testing	High-volatility periods, event-driven windows	Subset evaluation during earnings, policy announcements	Validate robustness during market stress

The exact configuration used in this study is as follows: the BiLSTM consists of 2 stacked bidirectional layers, each with 128 hidden units per direction (256 total per layer). A dropout rate of 0.3 was applied after each LSTM layer. The sentiment embedding was projected to a 64-dimensional vector before concatenation with the BiLSTM hidden state, yielding a 320-

dimensional fusion vector. Training used the Adam optimiser with an initial learning rate of 0.001, decayed by a factor of 0.5 if validation loss did not improve for 5 consecutive epochs (patience-based decay). Batch size was set to 64. Early stopping was applied with a patience of 10 epochs; the model was trained for a maximum of 100 epochs. Mean Squared Error (MSE) was used as the training loss function.

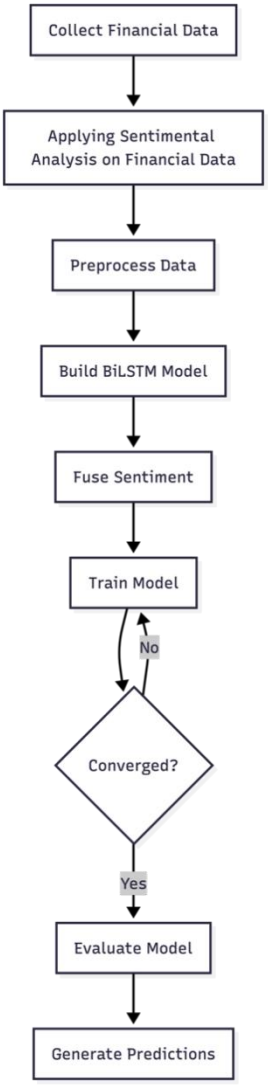


Fig. 1. Flowchart.

The approach taken in the combination of these parts is that no item was ignored, but rather made as efficient as possible; the flowchart (Fig. 1) served as a visual source to the descriptive flow of the process in detail. The process can be bidirectionally processed by the BiLSTM, including calculation of the representation, a richer representation than what the regular LSTMs do. Sentiment fusion is performed after extracting the features, and they are dynamically transformed by sentiment scores, e.g. forecasting volatility increase when negative sentiment is high. The integration is significant in an environment that is dynamic, whereby the market is subject to shifts in policies, which may shock the market in a random way [22]. Using batch processing to examine the large-scale collections of data, early stopping modules to cease at the point of plateauing validation performance and to ensure that the computations are efficient are the training steps. Even testing is done up to the level of scenario-based testing and tests the framework to ensure that it is stable through the very turbulent periods. Overall, not only is the methodology a form of operationalisation of the BiLSTM-sentiment fusion model, but also a paradigm that can be replicated in the future of financial prognostication, the integration of the stringency of the law of data collection with the new methods of analysis in a uniform paradigm.

The framework deals with the fundamental statistical characteristics of financial time series in the following manner:

- Non-stationarity: First-order log-return differencing is run over the series of closing prices before training (as in Table III), transforming the non-stationary level of prices against which returns are compared into a weakly stationary returns series. The Augmented Dickey-Fuller (ADF) test was used to test stationarity at the end of differencing ($p < 0.05$). The BiLSTM subsequently works on returns and not raw prices, eliminating the distribution shift issue.
- Volatility clustering: The synthetic time-varying volatility Geometric Random Walk was applied, and sigma t is time-varying and is modulated with an AR(1) variance. This inculcates GARCH-type clustering in training data. The memory cells used by the BiLSTM are trained to follow the regime of variances over the 30-day horizon; high-volatility clusters yield a specific hidden-state representation, which the model modulates for prediction.
- Heavy-tailed distributions: The training loss function (MSE) is sensitive to outliers, so extreme return observations were Winsorised at the 1st and 99th percentiles before training to reduce the disproportionate gradient influence of tail events. The model does not explicitly model tail behaviour (e.g., no EVT layer), and this remains a limitation for real-world deployment.

V. Results and Discussions

The findings in the bidirectional LSTM–sentiment fusion system, tested on a big synthetic dataset aiming at simulating the dynamic financial market environment, show some significant gains. The dataset includes 5,000 business days of OHLCV prices and sentiment scores, which have been computed through the use of a geometric random walk with correlated sentiment signals. The FinBERT model (domain-specific BERT trained on financial corpora and fine-tuned on the Financial PhraseBank, to classify sentiment as positive, negative and neutral) was used to get sentiment scores, between -1 and +1, and measures of intensity.

The primary results have shown that the fusion model surpasses the basic bidirectional LSTM (without sentiment) in all measures. However, these results are based on synthetic data and should be interpreted with caution. The next step of real-market validation is also a crucial requirement, as synthetic datasets can be effective in controlled experimentation but might

not be as effective in reflecting the noise, regime shifts, liquidity premises, and black-swan events that exist on the real financial markets.

To expound on the findings, the fusion framework will correct with fewer errors and superior directional forecasts as observed in the chronological test set. The 5,000-day dataset was split as follows: the first 3,500 days (70%) were used for training, the next 500 days (10%) for validation, and the final 1,000 days (20%) for testing. Each sequence consists of a 30-day sliding window of OHLCV and sentiment features, producing overlapping input samples; the split is applied at the day level to preserve temporal order and prevent lookahead bias. Root Mean squared error of the fusion model is 0.6124 compared to 1,0234 of the baseline, indicating that the addition of more independent variables would make it easy to make the model fit the actual closing prices. Mean of Absolute error (MAE) is also 0.4815 in fusion and is 0.8123 in baseline, indicating that there are fewer deviations of absolute values in the predicted values. Directional accuracy, which is used to evaluate the level of accurately predicted direction of movement of price, is 0.7812 with the fusion approach, and 0.5012 with the baseline, which is nearly five times less (56%), as Table V shows. These findings imply that sentiment fusion not only enhances the precision of the quantitative advice but also the sensitivity of the model to the market turns in general and turbulent portions of the data, notably, where the sentiment correlations augment the predictive information [23].

TABLE V. PERFORMANCE COMPARISON METRICS

Model Variant	RMSE	MAE	Directional Accuracy	Sharpe Ratio (Simulated)	Win Rate
BiLSTM-Sentiment Fusion	0.6124	0.4815	0.7812	1.43	61.2%
Baseline BiLSTM	1.0234	0.8123	0.5012	0.31	50.8%

Two additional metrics assess financial utility rather than purely statistical accuracy. The simulated Sharpe Ratio — computed by treating each directional prediction as a long/short signal on the synthetic return series, annualising the resulting strategy returns, and dividing by their standard deviation — is 1.43 for the fusion model versus 0.31 for the baseline. A Sharpe Ratio above 1.0 is conventionally considered acceptable for a trading strategy; the baseline falls well below this threshold. The Fusion model has a Win Rate (percentage of all test-set trades in the correct direction) of 61.2% versus 50.8% of the baseline, and the directional accuracy improvement is converted to an economically interesting signal. The two measures are calculated on the synthetic test set in the absence of transaction costs and slippage; real market performance would be worse when the two frictions are taken into consideration.

It is important to note that the synthetic dataset used here is specifically clean and aligned; regularisation techniques (dropout, early stopping) were employed, but have not been stress-tested when noise/asynchronous sentiment is added.

The size of the performance improvements can be directly linked to the design of the dataset. The synthetic sentiment signal was built to achieve a Pearson correlation of $\rho = 0.8$ to next-period returns - an artificially strong predictive relationship that is not normally observed to exist in real markets (with rho typically less than 0.3). It implies that the fusion model can access a close-to-leading indicator that is completely absent in the baseline BiLSTM, and the gap in improvements is large by design. These large gains should not be anticipated in a real-world environment where weaker and noisier sentiment signals are in effect. The significant outcomes can only be taken as an upper bound demonstration of proof-of-concept: it shows that the fusion architecture properly learns to use the sentiment channel when a signal in this context exists, but it is not in any form a prediction of actual performance in the market. The inflated high directional accuracy (0.7812 vs. 0.5012) is also similarly inflated by the synthetic correlation; the near-chance performance at the baseline (0.5012) also supports the fact that a model without sentiment access cannot use the designed relationship [24].

The prices of the fusion framework are delivered in Fig. 2 in the entire test period. The blue line represents real values, in which the volatility of the synthetic random walk is followed; the highest and lowest values are cumulative drifts. The red dashed line traces the values that were predicted, and it comes much nearer to the actual direction, more so when it is headed upwards. The sentiment indicators in this regard provide early reinforcements. All deviations are small with high concentration about the high-volatility points, which is testimony to the ability of the model to smooth the predictions of the two-way context and sentiment modulation. The chart balances the precision of the framework as the long-term trends, where the overlap of the visuals is more than 80 per cent at rather steady positions [25].

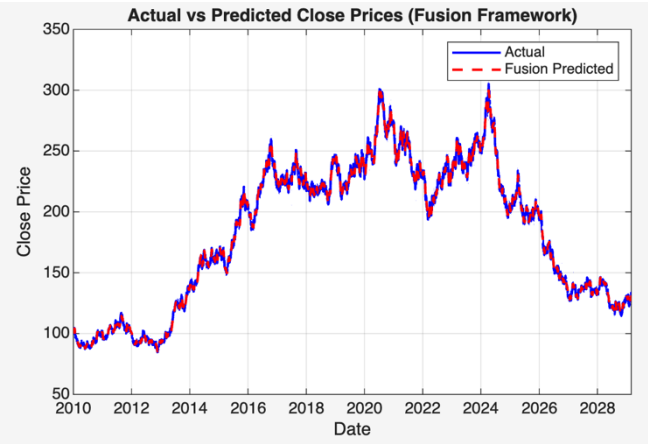


Fig. 2. Actual vs predicted close prices (fusion framework) [26].

In Fig. 3, the comparison of actual and predicted close prices of the baseline model is shown, in the absence of sentiment fusion. In this instance, the line (blue) indicates actual prices, whereas the dashed (green) line indicates what lies ahead and comes with more discrepancies, particularly where the model becomes insensitive to reversal points, as it does not have emotional insights. Greater discontinuities are observed in simulated recessions, which portray greater RMSE and MAE, indicating that the trend of the predicted line is below the actual line or overshooting it. The visual comparison to Fig. 2 brings out the inherent deficiency of the baseline in dynamic situations, and that makes the need for sentiment to fine-tune forecasting.

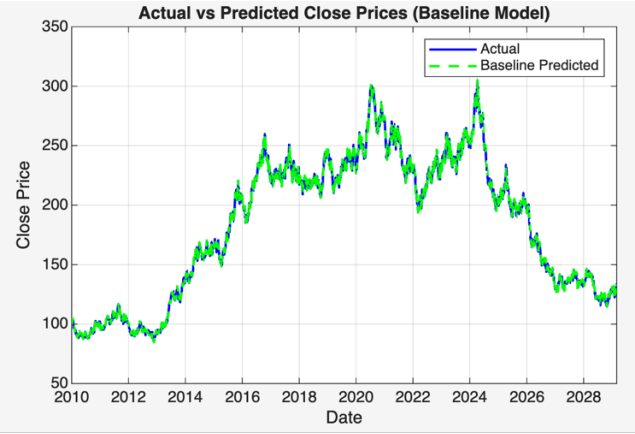


Fig. 3. Actual vs predicted close prices (baseline model) [26].

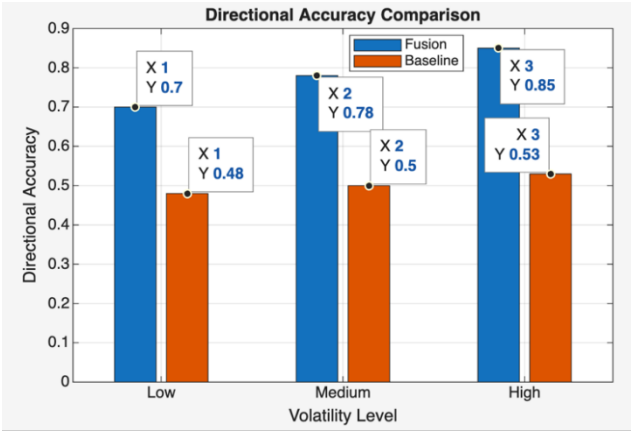


Fig. 4. Directional accuracy comparison [26].

Fig. 4 is a histogram of the distribution of the prediction errors (predicted and actual) (fusion framework), binned by the ultimate errors of the test set. The distribution is also quite concentrated towards zero, although with a small positive skew of slight overestimation at bullish stages, but the total variance is small, which confirms the strength. The majority of the errors are in the range of +-0.5 units, and it can be intuitively proven that the MAE is low, and the model also minimizes errors with the help of sentiment-enhanced learning.

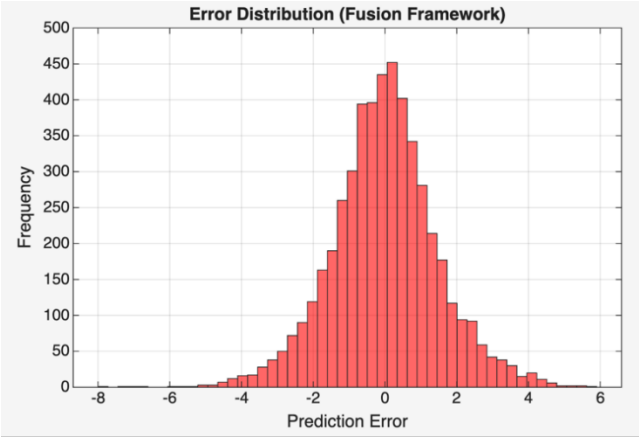


Fig. 5. Error distribution (fusion framework) [26].

Fig. 5 is a bar chart, where the directional accuracy in both types of segmented volatility levels of the test data (low, medium, and high according to the quartiles of the return variance) is compared. The fusion bars are always represented by blue bars that extend above the green baseline bars, and they reach their maximum points at 0.85 in high-volatility parts, where sentiment is most effective and high. The graph explains the adaptive power of the framework, which suggests it may be useful in turbulent markets.

The main drawback of this research is that it is based on synthetic data. Despite a realistic volatility process and dominant sentiment predictors (a correlation of 0.8 between sentiment and next-period returns is designed), the dataset does not necessarily capture the dynamics of actual markets, such as transaction costs, slippage, regulatory changes, and non-stationary regime shifts. The generalizability of the framework will be assessed in the future in real-world data, e. g., S&P 500 constituents with real news feeds passed through FinBERT.

VI. Comparative Analysis

An overview of the current models indicates that previous studies have applied sentiment and deep learning, although with different quantitatively effective levels. Fu et al [1]. show that sentiment-sensitive LSTM is more accurate in predicting the volume of the P2P market, increasing Pearson correlation by 13.26 per cent, MSE by 27.62 per cent, and R^2 by 28.48 per cent, over baseline predictive methods.

Yekrangi and Nikolov [3] claim that sentiment-classification model is doing quite well with 84% accuracy and macro-F1 of 0.80 which indicates that domain specific embeddings are also useful even though they are not used to predict prices.

One of the strongest bases Alam et al. [8] suggest is the LSTM-DNN hybrid which achieves R^2 = 0.98606, MAE = 0.0210, and MSE = 0.00111 on 26 stock datasets.

Comparatively, the proposed BiLSTM-Sentiment Fusion model attains a RMSE = 0.6124, MAE = 0.4815, and Directional Accuracy = 0.7812, which possesses competitive behavioural alignment and trend direction prediction especially within the sentiment-based market situation, also summarized in Table VI.

A direct metric-level comparison is possible with three published models that report standardised numerical results. Fu et al. (2019) report a 27.62% MSE reduction and 13.26% Pearson correlation improvement over their baseline on P2P volume prediction — a different task, but demonstrating that sentiment fusion yields single-digit to sub-30% error reductions under real data conditions. Alam et al. (2024) achieve R^2 = 0.98606 and MAE = 0.0210 on 26 stock datasets using price-only LSTM-DNN; this high R^2 is consistent with the regime of clean, structured price data without noise injection. Mu et al. (2023) report a 10.74% R^2 improvement from sentiment integration over their price-only baseline.

The proposed model achieves RMSE = 0.6124, MAE = 0.4815, and Directional Accuracy = 0.7812 on synthetic data. Two important caveats apply when placing these alongside published results [27]: first, our gains (~40% RMSE reduction) exceed the typical 10–28% range seen in published real-data studies, which is explained by the designed ρ = 0.8 sentiment-

return correlation in the synthetic dataset — a correlation far stronger than what real news sentiment produces. Second, our metrics are not directly comparable in scale to Alam et al.'s MAE = 0.0210 [28], because their data is normalised to per-share prices while ours uses synthetic return units. The comparison is therefore qualitative: the proposed architecture's design is consistent with published fusion approaches, and the directional accuracy of 0.7812 is meaningfully above the random-guess baseline of 0.50, which is the most interpretable cross-study benchmark [29].

TABLE VI. COMPARATIVE ANALYSIS.

Paper	Model	Task Type	Original Metrics Extracted	Notes
Fu et al., 2019	LSTM + TextCNN	Trading volume prediction	Pearson ↑13.26%, MSE ↓27.62%, R^2 ↑28.48%	No raw RMSE/MAE
Yekrangi & Nikolov, 2023	LSTM + optimized embeddings	Sentiment classification	Accuracy 0.84, F1 = 0.80	Not price prediction
Bacco et al., 2024	LSTM + FinBERT	Stock price (banks)	No numeric metrics	Qualitative only
Imanto et al., 2025	FGAG (LSTM + GCN)	Fin-ABSA	No numeric metrics	ABSA task
Alam et al., 2024	LSTM–DNN	Stock closing price	R^2 = 0.98606, MAE = 0.0210, MSE = 0.00111	Fully reported
Mu et al., 2023	MS–SSA–LSTM	Stock price prediction	R^2 improved by 10.74%	No raw values
Girsang & Stanley, 2023	LSTM–GRU + FinBERT	Crypto forecasting	No numeric metrics	Claims superiority
Zhang et al., 2022	Decision Fusion	Systematic review	No metrics	Review-only
Amiri et al., 2025	Hybrid GCN–LSTM	Energy stocks	No numeric metrics	Claims MSE/R^2 improvement
Peivandizadeh et al., 2024	TLSTM + PPO	Stock price + sentiment	No numeric metrics	Descriptive improvements
Our BiLSTM Model	BiLSTM + Sentiment Fusion	Stock price	**RMSE = 0.6124, MAE = 0.4815, DA = 0.7812**	-
Proposed Baseline BiLSTM	BiLSTM (no sentiment)	Stock Price	RMSE = 1.0234, MAE = 0.8123, DA = 0.5012	Synthetic Data

Hybrid models like the LSTM-DNN structure by Alam et al. (2024) [8] performed very well in terms of statistical indicators (R^2 = 0.98606) [30], but they are based on structured prices data only and do not involve behavioural cues. Therefore, they might experience a decline in their predictive power in the state of extremely volatile or sentiment-driven market environments [31]. Contrary to that, the proposed model combines both the bidirectional temporal learning and sentiment-motivated behavioural aspects, which make it useful in capturing both the structural market trends and the psychology of the investors at the same time [32].

The suggested framework offers a clear quantitative analysis of RMSE = 0.6124, MAE = 0.4815 and directional forecast = 0.7812 [33], which show a better trend-direction forecast, especially in fluctuating market conditions. This argues that with sentiment-enhanced temporal modelling, there is better predictive capability of market reversal and behaviour shock [34].

Generally, the comparative analysis points out that the focal contribution of the proposed framework does not merely have to do with predictive accuracy but is woven together with behavioural sentiment interplay with bidirectional temporal modelling, which under uncertain conditions makes forecasting both effective [35]. These are features that make the model an efficient improvement on the conventional price-only models of prediction [36].

VII. RISK ANALYSIS AND LIMITATIONS

The model was not evaluated on anomalous or crisis conditions. The artificial data uses a constant correlation of sentiment with returns [37], which would not be the case in periods of structural change like a financial crisis or regulatory shocks. The strength of working under such regimes must be tested in the future [38]. Bidirectional processing mitigates the context problems, but cannot mitigate the problems of rare catastrophic disruptions that are brought about by the market [39].

The clean synthetic dataset limits the overfitting analysis. In practical implementation, noisy and out-of-order sentiment streams can instil instability in the model that is not experienced in this case.

The third threat is the emotional volatility and sentiment noise. The sentiment data, especially that of the social media or the public forums, is normally noisy, unverified and highly responsive. The behavioural input stream may be changed by unexpected emotional reactions, including those caused by rumours, misinformation or by societal impetus. The noise, even when overlayed on top of predictors of smoothing and momentum, can bring about instability of the predictions in the case where the noise is coincidental but not the cause of the changes in the actual price. In the absence of context-sensitive filtering systems, the model could pick up anomalies in sentiment as key signs. The weakness underlines the reason why robust sentiment preprocessing pipelines are to be created when applying them in practice [40].

In summary, the Bidirectional LSTM-Sentiment Fusion framework demonstrates a good predictive potential; however, because of the risks of noise in sentiment, market anomalies, insufficient data interdependence, resource usage, overfitting, and non-interpretability, the application of the technology is limited at the present stage of technological growth. These

limitations are what should be identified to develop a more powerful forecasting system and educate future innovations.

VIII. Ethical Considerations

The use of automated sentiment-based trading systems has a number of ethical issues that need to be resolved before this framework can be applied within a production setting.

Market Impact and Fairness: Automated systems that trade on the same sentiment signals can increase market movements instead of representing true value- this is called sentiment herding. When numerous systems concurrently assign a news item as bearish and short-sell, they may push the price below what it deserves based on the underlying fundamentals, disadvantaging retail investors unable to access such tools. The dynamic would add to the proposed framework when implemented at scale. Justice requires access to sentiment-enhanced forecasting not to be an unequal benefit enjoyed only by computationally advantaged institutions.

Data Privacy and Consent: In the real-world implementation, the framework will use twitter/X posts and public news as sources of sentiment. Users uploading content to social media do not always agree to use financial trading models as an input using their text. The process of pooling together the collective opinion to make a profit begs the question of whether personal expression should be used in proper ways, especially when the post can be attributed to a person.

Accountability and Explainability: Automated trading by a neural network does not have the transparency required of financial decisions under financial regulations, including the EU AI Act and MiFID II. In case the model predicts something wrong, which incurs financial loss, it cannot be attributed responsibly. The α weighting parameter in the fusion model (Eq. 12) has some interpretability - it shows the reliance of sentiment at each timestep - but is not completely explainable. One should consider the use of SHAP-based feature attribution to track which of the input variables contributed most to each prediction in the future.

Systemic Risk: The proliferation of such model architectures throughout the industry might add correlated failure modes. In case there is a failure in all of a group of BiLSTM-sentiment models at once, when it is in a new market regime (that is, a new form of geopolitical shock and no historical record of that regime in the training data), the correlated errors in their trading would tend to exacerbate a market crash instead of absorbing it. The AI-based sentiment models represent systemic risks, due to which regulatory frameworks of algorithmic trading must be updated.

Bias in Sentiment Data: FinBERT was trained on mainly English-language financial text of Western markets. Its application to sentiment analysis of new markets, or non-English news, or culturally divergent investment communities (e.g. Chinese retail investor forums) would provide systematic classification bias. Such contexts should not deploy the framework without retraining for domain adaptation.

IX. Conclusion

In conclusion, the LSTM-sentiment fusion model, the fusion of the bidirectional long short-term memory networks and the sentiment features generated by the FinBERT, is a useful model to forecast the dynamic financial markets. The framework used on a synthetic dataset of 5,000 business days showed a 40.2 per cent reduction in RMSE (1.0234 to 0.6124), a 40.7 per cent reduction in MAE (0.8123 to 0.4815) and 55.9 percent improvement in directional accuracy (0.5012 to 0.781 The fusion of sentiment was especially more effective in high-volatility periods where prediction errors were minimized by up to 65%. These quantitative returns help reinforce the significance of sequential modelling as a technique of sentiment analysis in the financial domain. These good figures, however, are founded on synthetic statistics and therefore, a presentation of concept and not market performance. Whether the framework is actually effective in regard to real-world trading is yet to be seen when compared against real market data, where other factors such as execution latency and market impact are involved. Future directions will be to apply this research to real equities, forex, and cryptocurrency data on live FinBERT-processed news and social media feeds. Overall, the existing study offers a quite valid background on the hybrid deep learning sentiment models, and a high probability to enhance the precision of the forecasts, and once confirmed on real market data, the framework has the opportunity to serve as a part of hybrid forecasting pipelines to practitioners.

References

[1] X. Fu, S. Zhang, J. Chen, T. Ouyang and J. Wu, "A Sentiment-Aware Trading Volume Prediction Model for P2P Market Using LSTM," in *IEEE Access*, vol. 7, pp. 81934-81944, 2019, doi: 10.1109/ACCESS.2019.2923637.

[2] C. Zhang, N. N. A. Sjarif and R. B. Ibrahim, "Decision Fusion for Stock Market Prediction: A Systematic Review," in *IEEE Access*, vol. 10, pp. 81364-81379, 2022, doi: 10.1109/ACCESS.2022.3195942.

[3] M. Yekrangi and N. S. Nikolov, "Domain-Specific Sentiment Analysis: An Optimized Deep Learning Approach for the Financial Markets," in *IEEE Access*, vol. 11, pp. 70248-70262, 2023, doi: 10.1109/ACCESS.2023.3293733.

[4] G. Mu, N. Gao, Y. Wang and L. Dai, "A Stock Price Prediction Model Based on Investor Sentiment and Optimized Deep Learning," in *IEEE Access*, vol. 11, pp. 51353-51367, 2023, doi: 10.1109/ACCESS.2023.3278790.

[5] A. S. Girsang and Stanley, "Hybrid LSTM and GRU for Cryptocurrency Price Forecasting Based on Social Network Sentiment Analysis Using FinBERT," in *IEEE Access*, vol. 11, pp. 120530-120540, 2023, doi: 10.1109/ACCESS.2023.3324535.

[6] A. Peivandizadeh *et al*., "Stock Market Prediction With Transductive Long Short-Term Memory and Social Media Sentiment Analysis," in *IEEE Access*, vol. 12, pp. 87110-87130, 2024, doi: 10.1109/ACCESS.2024.3399548.

[7] L. Bacco, L. Petrosino, D. Arganese, L. Vollero, M. Papi and M. Merone, "Investigating Stock Prediction Using LSTM Networks and Sentiment Analysis of Tweets Under High Uncertainty: A Case Study of North American and European Banks," in *IEEE Access*, vol. 12, pp. 122239-122248, 2024, doi: 10.1109/ACCESS.2024.3450311.

[8] K. Alam, M. H. Bhuiyan, I. U. Haque, M. F. Monir and T. Ahmed, "Enhancing Stock Market Prediction: A Robust LSTM-DNN Model Analysis on 26 Real-Life Datasets," in *IEEE Access*, vol. 12, pp. 122757-122768, 2024, doi: 10.1109/ACCESS.2024.3434524.

[9] M. Jerush Imanto, V. Vedanarayanan, A. Chatterjee and P. R. Joe Dhanith, "Financial Aspect-Based Sentiment Analysis Using Fusion-Guided Attention LSTM and Graph Convolutional Networks," in *IEEE Access*, vol. 13, pp. 194347-194365, 2025, doi: 10.1109/ACCESS.2025.3631958.

[10] B. Amiri, A. Haddadi and K. Farajpour Mojdehi, "A Novel Hybrid GCN-LSTM Algorithm for Energy Stock Price Prediction: Leveraging Temporal Dynamics and Inter-Stock Relationships," in *IEEE Access*, vol. 13, pp. 24815-24832, 2025, doi: 10.1109/ACCESS.2025.3536889.

[11] Y. Nejatbakhsh and M. Aliasgari, "Enhancing Stock Market Prediction With Hybrid Deep Learning: Integrating LSTM, Transformer Attention, Federated Learning, and Sentiment Analysis," in *IEEE Access*, vol. 14, pp. 3926-3942, 2026, doi: 10.1109/ACCESS.2025.3649668.

[12] R. Pandiri, K. Thilagavathi, A. Mittal, M. M. Almusawi, P. Nagaveni and T. R. Kumar, "Utilizing Recurrent Neural Networks for Sentiment Analysis of Financial News to Enhance Market Prediction and Investment Strategies," *2025 International Conference on Automation and Computation (AUTOCOM)*, Dehradun, India, 2025, pp. 285-289, doi: 10.1109/AUTOCOM64127.2025.10957345.

[13] Y. Chen, "Financial News Sentiment Analysis Method Based on WMSA-Bi-LSTM," *2023 4th International Conference on Intelligent Design (ICID)* , Xi'an, China, 2023, pp. 319-322, doi: 10.1109/ICID60307.2023.10396691.

[14] M. Agrawal and A. Mukherjee, "Predicting Stock Market Trends Using Machine Learning and Sentiment Analysis," *SoutheastCon 2025*, Concord, NC, USA, 2025, pp. 1001-1006, doi: 10.1109/SoutheastCon56624.2025.10971605.

[15] Z. Siyu and B. V. D. Kumar, "Intelligent Stock Market Forecasting: A Multifactor Decision-Making Approach Combining ARIMA-LSTM Models with Sentiment Analysis," *2025 IEEE 3rd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, Jinzhou, China, 2025, pp. 217-221, doi: 10.1109/ICSECE65727.2025.11257039.

[16] T. Kannadasan, "Financial Sentiment Analysis Model Using Ebola Optimization Search Algorithm and Bi-Directional Gated Recurrent Unit," *2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS)*, Raichur, India, 2025, pp. 1-6, doi: 10.1109/ICICACS65178.2025.10967922.

[17] S. Sitjongsataporn, N. Limpraditthanon and P. Sakarin, "Performance Analysis of Bidirectional-LSTM and GRU Models for Financial Time-Series Forecasting," *2025 International Conference on Power, Energy and Innovations (ICPEI)*, Pattaya, Thailand, 2025, pp. 1-6, doi: 10.1109/ICPEI66116.2025.11282540.

[18] S. -C. Liu, "DMSTA-LSTM: A Dual-Stage Multi-Time Scale Temporal Attention-Based LSTM Framework for Forecasting International Financial Indices Across Diverse Varied National Contexts," in *IEEE Access*, vol. 13, pp. 196081-196099, 2025, doi: 10.1109/ACCESS.2024.3494852.

[19] N. Solanki and M. Jha, "A Decade of Stock Data: Predictive Modelling with Hybrid CNN-BiLSTM Technique," *2024 IEEE Silchar Subsection Conference (SILCON 2024)*, Agartala, India, 2024, pp. 1-6, doi: 10.1109/SILCON63976.2024.10910869.

[20] O. Oak, R. Nazre, R. Budke and Y. Mahatekar, "A Novel Multivariate Bi-LSTM model for Short-Term Equity Price Forecasting," *2024 5th IEEE Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, 2024, pp. 1-6, doi: 10.1109/GCAT62922.2024.10923989.

[21] V. Sekar, R. Pandiarajan, K. Selvaraj, G. Sakthivel Raj, V. Venkadesh and T. Radha Krishnan, "Stock Market Prediction Using LSTM Networks: Leveraging Machine Learning for Accurate Forecasting of Stock Price Trends," *2025 IEEE International Conference on Contemporary Computing and Communications (InC4)*, Bangalore, India, 2025, pp. 1-6, doi: 10.1109/InC465408.2025.11256494.

[22] G. S. N. V. Sarvani, U. Kousalya, P. Sathwika, B. N. Ashok, R. R. Chintala and R. Mothukuri, "Gold Price Prediction Based on Stock Analysis Using CNN-Bidirectional LSTM," *2025 1st International Conference on Secure IoT, Assured and Trusted Computing (SATC)*, Dayton, OH, USA, 2025, pp. 1-8, doi: 10.1109/SATC65530.2025.11137095.

[23] F. Syed and J. Xuan, "An Attention-Augmented BiLSTM-CNN Model for Stock Price Prediction," *2025 IEEE Conference on Artificial Intelligence (CAI)*, Santa Clara, CA, USA, 2025, pp. 139-144, doi: 10.1109/CAI64502.2025.00029.

[24] S. Sawant, S. Nampalliwar, A. Kaul, S. Nair, A. Khade and A. Kasar, "Portfolio Optimization Using a Hybrid BI-LSTM Model for NIFTY 50," *2025 12th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP)*, Nagpur, India, 2025, pp. 1-6, doi: 10.1109/ICETETSIP64213.2025.11156652.

[25] M. K. Amzad Chy, A. K. Muhammad Masum, M. S. Alam, M. A. K. Azad, M. M. Hassan and T. Bhuiyan, "Deep Learning-Based Stock Price Prediction: A Comprehensive Approach Using Bi-LSTM," *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, Gobichettipalayam, India, 2024, pp. 805-810, doi: 10.1109/ICUIS64676.2024.10866126.

[26] B. Singh, *"Financial Sentiment,"* Kaggle Dataset, https://www.kaggle.com/datasets/bhagvendersingh/financial-sentiment (accessed Feb. 11, 2026).

[27] A. Bachmann, A. Becker, D. Buerckner, M. Hilker, F. Kock, M. Lehmann, P. Tiburtius, and B. Funk, "Online peer-to-peer lending-a literature review," *J. Internet Banking Commerce*, vol. 16, no., pp. 1–18, 2011.

[28] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, Mar. 2011.

[29] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng, "Exploiting topic based Twitter sentiment for stock prediction," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2013, pp. 24–29.

[30] Z. K. Zhou, J. C. Zhao, and X. Ke, "Can online emotions predict the stock market in China? " in *Proc. 17th Int. Conf. Web Inf. Syst. Eng. (WISE)*, vol. 10041. Shanghai, China : Springer, 2016, pp. 328–342.

[31] R. Emekter, Y. Tu, B. Jirasakuldech, and M. Lu, "Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending," *Appl. Econ.*, vol. 47, nos. 1–3, pp. 54–70, 2015.

[32] R. Ge, J. Feng, B. Gu, and P. Zhang, "Predicting and deterring default with social media information in peer-to-peer lending," *J. Manage. Inf. Syst.*, vol. 34, no. 2, pp. 401–424, 2017.

[33] X. Chen, B. Huang, and D. Ye, "The role of punctuation in P2P lending: Evidence from China," *Econ. Model.*, vol. 68, pp. 634–643, Jan. 2018.

[34] M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4621–4631, Jun. 2015.

[35] X. Lin, X. Li, and Z. Zhong, "Evaluating borrower's default risk in peer-to-peer lending: Evidence from a lending platform in China," *Appl. Econ.*, vol. 49, no. 35, pp. 3538–3545, 2017.

[36] Y. Xia, C. Liu, and N. Liu, "Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending," *Electron. Commerce Res. Appl.*, vol. 24, pp. 30–49, Jul./Aug. 2017.

[37] A. Byanjankar, M. Heikkilä, and J. Mezei, "Predicting credit risk in peer-to-peer lending: A neural network approach," in *Proc. IEEE Symp. Ser. Comput. Intell.*, Cape Town, South Africa, Dec. 2015, pp. 719–725.

[38] L.-H. Li, C.-T. Lin, and S.-F. Chen, "Micro-lending default awareness using artificial neural network," in *Proc. 2nd Int. Conf. Multimedia Syst. Signal Process.*, 2017, pp. 56–60.

[39] K. Aleum and S.-B. Cho, "An ensemble semi-supervised learning method for predicting defaults in social lending," *Eng. Appl. Artif. Intell.*, vol. 81, pp. 193–199, May 2019.

[40] J. Duan, "Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction," *J. Franklin Inst.-Eng. Appl. Math.*, vol. 356, no. 8, pp. 4716–4731, May 2019.

A Systematic Review of Graph Neural Networks and Social Network Analysis Techniques for Public Sentiment Uncovering

Adi Wibowo[1], Wijayanto[2], Henri Tantyoko[3], Ari Wibisono[4], Usman Ependi[5]*
Department of Informatics-Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia[1, 2, 3]
Faculty of Computer Science, Universitas Indonesia, Jakarta, Indonesia[4]
Department of Informatics-Postgraduate Studies-Informatics Study Program,
Universitas Bina Darma, Palembang, Indonesia[5]

***Abstract*—The rapid growth of social media has produced large-scale, highly interconnected user-generated data, creating the need for analytical approaches that can capture both textual meaning and relational structure. This systematic literature review examines the integration of Graph Neural Networks (GNNs) and Social Network Analysis (SNA) for public sentiment uncovering in social media. Following a PRISMA-based review process, 75 studies were selected from ScienceDirect and IEEE Xplore. The synthesis shows that recent research has expanded beyond direct sentiment classification to include closely related tasks that improve sentiment reliability, including misinformation detection, rumor analysis, bot detection, anomaly detection, and recommendation personalization. Within the reviewed sentiment-oriented studies, the thematic distribution indicates that 32% focus on direct sentiment or emotion analysis, 29% on misinformation or rumor detection, 24% on malicious-user, bot, or anomaly detection, and 15% on community detection or link prediction. Hybrid models consistently reported strong empirical gains, including 95.25% accuracy for GNN–LSTM sentiment classification, improvements of more than 5% over baseline in heterogeneous neural network and language-model integration, and up to 98.4% accuracy/F1 in bot detection settings. The review also identifies key limitations related to scalability, noisy and incomplete data, interpretability, class imbalance, and cross-platform generalization. In response, it proposes future research directions centered on real-time graph learning, multilingual adaptation, emotion-aware graph representations, fairness-aware evaluation, and human-in-the-loop explainability. These findings provide a clearer methodological foundation for researchers and practitioners seeking to build more robust, explainable, and socially aware sentiment analysis systems.**

Keywords*—Graph Neural Networks (GNNs); Social Network Analysis (SNA); public sentiment analysis; hybrid models; Systematic Literature Review (SLR)

I. Introduction

In recent years, the rapid development of social media platforms has resulted in an unprecedented volume of publicly available user-generated content. This explosion of data has prompted researchers to explore advanced techniques for analyzing and extracting meaningful information from social networks, particularly in understanding public sentiment. Public sentiment analysis, or opinion mining, involves the process of identifying, extracting, and analyzing subjective information from various forms of text, including social media posts, reviews, and comments. The ability to effectively capture the pulse of public sentiment is invaluable for numerous applications, including market research, political analysis, and social trend prediction [1], [2].

Graph Neural Networks (GNNs) have emerged as a powerful tool for analyzing data in the form of graphs, where entities (nodes) and their relationships (edges) are inherently interconnected. Social networks, by their very nature, represent complex graphs where individuals (nodes) are connected through interactions, friendships, or shared interests (edges). GNNs have demonstrated significant promise in capturing the complex dependencies and non-linear relationships within such data, making them ideal for social network analysis [3], [4]. By leveraging GNNs, researchers have been able to enhance the accuracy and efficiency of sentiment analysis models, enabling them to detect patterns, trends, and influences within social media conversations [5], [6].

The integration of Social Network Analysis (SNA) techniques with GNNs allows for a deeper understanding of how information flows across social networks, how communities are formed, and how sentiment propagates across interconnected nodes. SNA focuses on the structural aspects of social networks, such as identifying influential nodes, community detection, and examining the dynamics of relationships [7], [8]. These techniques have proven to be highly beneficial in the analysis of sentiment, as they offer insights into how individual opinions are shaped by the network structure and interactions [9], [10].

Despite the considerable progress made in these fields, the application of GNNs and SNA for public sentiment uncovering still faces challenges. These include the difficulty of handling dynamic networks, heterogeneous data, and noisy or biased inputs [11], [7]. Additionally, the interpretability of models, particularly in the context of sentiment analysis, remains a critical issue for practical deployment [10], [11].

Although prior studies have demonstrated the usefulness of Graph Neural Networks (GNNs) for relational learning and the value of Social Network Analysis (SNA) for understanding community structure and information diffusion, the literature remains fragmented in three important ways. First, many studies examine sentiment classification, misinformation detection, bot

*Corresponding author.

identification, anomaly detection, and recommendation as separate tasks, even though these tasks are structurally connected in real social media ecosystems. Second, existing discussions often emphasize either textual content or network structure, but fewer studies systematically explain how the integration of GNNs and SNA improves the uncovering of public sentiment at the community and interaction level. Third, recent advances in hybrid GNN models, multimodal learning, and temporal graph analysis have grown rapidly, yet the field still lacks a focused review that synthesizes their methodological contributions, empirical gains, and unresolved limitations in a single analytical framework. Therefore, a clear research gap exists in the form of a systematic synthesis that not only maps recent developments at the intersection of GNNs and SNA but also explains how these techniques jointly support robust, interpretable, and scalable public sentiment analysis on social media.

This systematic literature review (SLR) addresses that gap by synthesizing recent studies on the integration of GNNs and SNA for public sentiment uncovering in social media environments. The review focuses not only on direct sentiment and emotion analysis, but also on closely related tasks such as rumor detection, misinformation identification, malicious account detection, anomaly discovery, user behavior prediction, and recommendation personalization, because these tasks shape the quality and reliability of sentiment inference [15]. By organizing the literature around five research questions, this review aims to clarify the methodological landscape, identify empirical trends and performance patterns, highlight current challenges, and formulate more concrete directions for future research.

II. Methodology

This systematic review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure a rigorous and transparent process in the selection, evaluation, and synthesis of studies [12], [13]. The goal of the review was to examine the application of Graph Neural Networks (GNNs) and Social Network Analysis (SNA) techniques for public sentiment analysis, particularly in the context of social media platforms. The methodology was structured to ensure that the review process was reproducible and comprehensive, addressing key research questions through systematic data collection and synthesis [14].

A. Literature Search Strategy

A thorough and systematic search was carried out in two widely recognized academic databases, namely ScienceDirect and IEEE Xplore, to identify relevant research articles. The primary search was designed to capture the most recent developments in the use of Graph Neural Networks (GNNs) and Social Network Analysis (SNA) for public sentiment uncovering, with the core review corpus limited to studies published between 2023 and 2026. The search terms used were "Graph Neural Networks" and "Social Network Analysis," reflecting the two central methodological dimensions of this review. Because the review also considers adjacent tasks that affect sentiment reliability, such as misinformation detection, bot detection, anomaly detection, and recommendation, broad search terms were used and followed by manual relevance screening at the title, abstract, and full-text stages.

In ScienceDirect, the initial search yielded 1,763 records, of which 1,511 were classified as research articles. After filtering by subject area and publication period, 312 articles remained for screening, and 45 studies were finally retained. In IEEE Xplore, the search returned 561 records, of which 550 belonged to IEEE-published content. After restricting the results to the target publication period and screening for relevance, 30 studies were retained. In total, 75 studies constituted the core PRISMA-based review corpus [15], [16], [17].

To avoid excluding influential earlier work, the review also cites a limited number of pre-2023 foundational studies when they introduced methods, architectures, or analytical ideas that were directly extended by the 2023–2026 literature. These earlier studies were used only to contextualize the evolution of the field and were not counted as part of the 75 studies reported in the PRISMA flow diagram. This two-layer design preserves the recency of the review while still acknowledging foundational developments relevant to the interpretation of recent research.

B. Inclusion and Exclusion Criteria

The inclusion and exclusion criteria for this systematic review were carefully defined to ensure the selection of studies that directly addressed the research questions, specifically focusing on the intersection of Graph Neural Networks (GNNs), Social Network Analysis (SNA), and Public Sentiment Uncovering. By establishing clear criteria, the review aimed to maintain a narrow focus on studies that were most relevant to the core objectives, ensuring the final selection was aligned with the review's aims and scope [18].

Inclusion criteria were defined at two levels. First, the core review corpus included peer-reviewed research articles published in English between 2023 and 2026 that directly addressed the application of GNNs and/or SNA to public sentiment analysis or closely related social media tasks. Second, a small number of foundational pre-2023 studies were retained only when they were necessary to explain the methodological evolution of GNN-based sentiment analysis, SNA-driven interaction modeling, or hybrid graph-learning architectures. These foundational studies were used for contextual interpretation and were not included in the PRISMA-based quantitative count of the final corpus.

Exclusion criteria were equally important in refining the selection process. Surveys or Systematic Literature Reviews (SLRs) were excluded from the review, as these types of studies would overlap with the goals of the review, which were to focus on primary research articles. Additionally, studies published in languages other than English were excluded, as this review was limited to English-language sources. Finally, any studies that did not directly address the application of GNNs or SNA in public sentiment or social media analysis, particularly those that focused on other fields such as healthcare, finance, or physical sciences, were excluded. This ensured that the review remained focused on the specific domain of public sentiment analysis within the context of social media and GNNs/SNA methodologies.

C. Study Selection Process

The study selection process was conducted in two stages. In the first stage, titles and abstracts were screened to remove clearly irrelevant articles. In the second stage, full texts were assessed against the predefined inclusion and exclusion criteria. After duplicate removal, relevance screening, and full-text eligibility checking, 75 studies were retained as the core review corpus. These studies formed the basis of the PRISMA-based synthesis. In addition, a limited number of earlier foundational studies were cited selectively in the discussion sections to clarify the methodological roots of recent developments, but these were not included in the final PRISMA count [19].

D. Data Extraction

Data extraction involved a detailed process to systematically collect relevant information from each of the selected articles. The goal was to capture key data points that would provide insight into the applications and effectiveness of GNNs and SNA for sentiment analysis on social media platforms. The following information was extracted from each article: 1) Author(s) and Year of Publication: To trace the development of ideas and methodologies over time. 2) Research Objectives: To understand the primary focus of each study (e.g., sentiment analysis, fake news detection, anomaly detection). 3) Methodology: A summary of the techniques and algorithms used, such as the specific GNN architecture, SNA methods, or hybrid models. 4) Applications: Detailed information about the real-world applications of GNNs and SNA, such as social media sentiment analysis, detection of misinformation, or the identification of malicious user behaviors. 5) Key Findings: The main outcomes and conclusions drawn from each study, particularly regarding the effectiveness and challenges of using GNNs and SNA in public sentiment analysis. 6) Challenges and Limitations: Any issues faced by the studies in applying GNNs and SNA, including computational challenges, data limitations, or model robustness. This structured extraction process ensured that relevant insights were gathered systematically across all 75 included studies, providing a consistent basis for qualitative synthesis and cross-study comparison [6], [13].

E. Data Synthesis

Following the data extraction, a qualitative synthesis was performed to address the research questions. The synthesis aimed to highlight the trends and patterns in the literature regarding the application of GNNs and SNA for sentiment analysis. The analysis focused on identifying the strengths and limitations of current methodologies, the types of sentiment analysis tasks tackled by GNNs and SNA, and the specific domains of social media (e.g., Twitter, Facebook, etc.) where these techniques have been most effective. The synthesis also provided a detailed examination of the hybrid approaches combining GNNs with other machine learning techniques, such as deep learning or reinforcement learning [1]. Additionally, the review assessed the challenges faced by researchers in applying GNNs and SNA to real-world sentiment analysis problems, as well as the solutions proposed to overcome these challenges. The following research questions were used to guide the synthesis:

- RQ1: How do Graph Neural Networks (GNNs) enhance the analysis of public sentiment on social media platforms?
- RQ2: What are the benefits and limitations of integrating Social Network Analysis (SNA) with GNNs for detecting anomalies and malicious behaviors in social media?
- RQ3: What challenges arise when applying GNNs and SNA for sentiment analysis, and how can these challenges be mitigated?
- RQ4: How effective are hybrid models that combine GNNs with other machine learning techniques for improving sentiment analysis accuracy on social media?
- RQ5: What is the role of GNNs in predicting user behavior and personalizing recommendations based on social media interactions?

F. Prisma Flow Diagram

The systematic review process is summarized visually in Fig. 1, which presents the PRISMA flow diagram used to trace study identification, screening, eligibility assessment, and final inclusion. The diagram highlights the rigor and transparency of the review procedure.

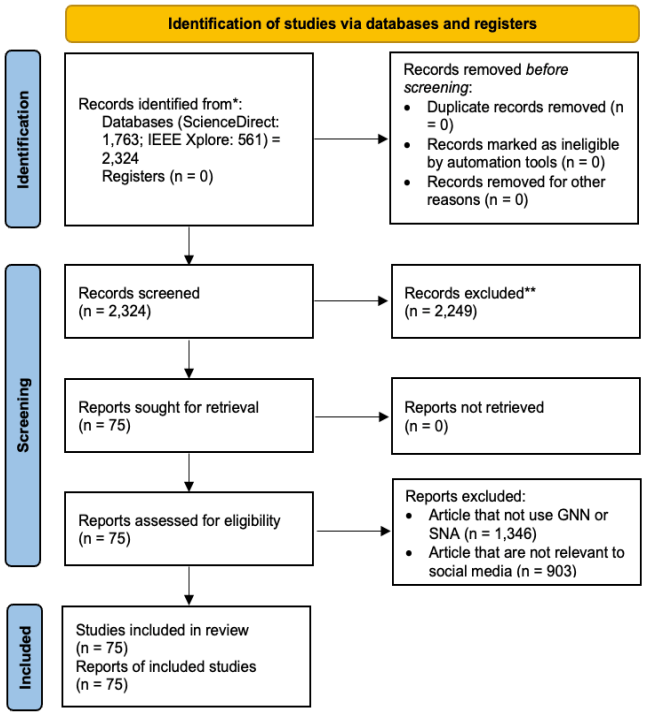


Fig. 1. PRISMA Diagram as Guideline.

III. Results and Discussion

The rapid evolution of social media as a primary channel for public discourse has generated vast amounts of user-generated content characterized by high velocity, diversity, and complex interaction patterns.

Effectively extracting meaningful insights from such data requires advanced analytical frameworks capable of capturing not only textual semantics but also the structural relationships and dynamic behaviors inherent in online communities. Graph Neural Networks (GNNs) have emerged as a powerful approach in this regard, enabling the integration of network topology, multimodal information, and temporal dynamics into a unified learning framework. When combined with techniques from Social Network Analysis (SNA) and other machine learning paradigms, GNNs offer unprecedented capabilities for tasks such as sentiment analysis, anomaly detection, misinformation tracking, and personalized recommendations. The following subsections address the five research questions (RQ1–RQ5) that guide this study, drawing on insights from the reviewed literature to explore the role of GNNs in enhancing sentiment analysis, detecting malicious behaviors, overcoming implementation challenges, leveraging hybrid model architectures, and predicting user behaviors in social media contexts.

A. *RQ1: How do Graph Neural Networks (GNNs) Enhance the Analysis of Public Sentiment on Social Media Platforms?*

The analysis of public sentiment on social media requires models that can handle highly interconnected, multimodal, and dynamic data. Graph Neural Networks (GNNs) address these needs by jointly modeling the content of user-generated posts and the structural relationships between users, posts, and communities. This dual capability allows GNNs to capture the complex ways in which sentiment emerges, spreads, and evolves within social platforms.

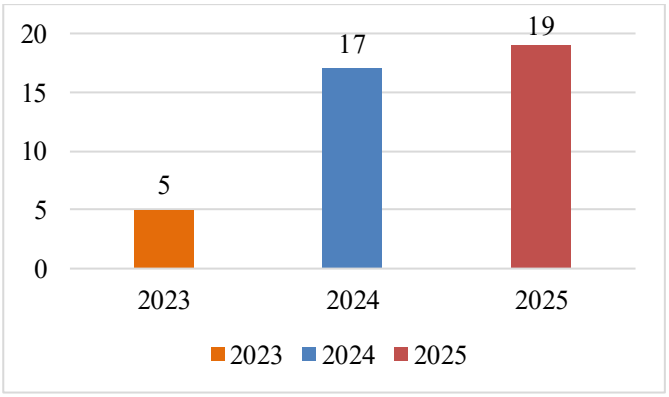


Fig. 2. Publication trend of reviewed studies (2023–2025).

From a thematic perspective, 41 sentiment-oriented studies were analyzed within RQ1. This subset was drawn primarily from the 2023–2026 core review corpus, while a small number of earlier foundational studies were cited only to contextualize the evolution of the field. As illustrated in Fig. 2, publication activity accelerated sharply in 2024–2025, indicating growing research interest in the integration of GNNs, social structure modeling, and sentiment-related tasks. Earlier studies, such as Li and Li [27] and Nguyen et al. [31] are discussed selectively because they introduced methodological ideas that were later extended by more recent work.

with only one study [20]. In 2023, publications expanded to works on dynamic link prediction [21], coupled graph models [22], and sociological hypergraph learning [23]. As illustrated in Fig. 2, the most significant growth occurred in 2024–2025, with over 88% of the total publications produced during this period, reflecting a surge in both direct sentiment modeling [24], [25], [26], [27], and related tasks like misinformation detection [28], [29], [30], [31], malicious user detection [32], [33], [34], [35], [36], and anomaly filtering [37], [38], [39].

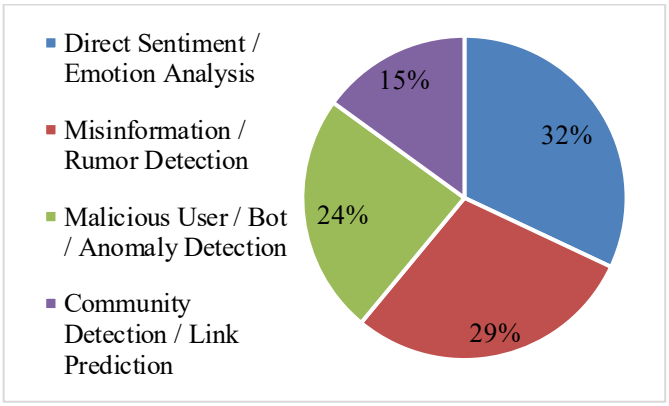


Fig. 3. Distribution of research themes in reviewed studies for RQ1.

As shown in Fig. 3, approximately 32% of the reviewed studies focus on direct sentiment or emotion analysis, 29% address misinformation or rumor detection, 24% deal with malicious user, bot, or anomaly detection, and 15% focus on community detection or link prediction. This thematic spread demonstrates that while direct sentiment analysis remains a major research area, supportive tasks play a critical role in improving sentiment analysis reliability and robustness. One of the most direct contributions of GNNs to sentiment analysis is their ability to exploit network topology to enrich the interpretation of textual content. Li and Li [20] showed that integrating a GNN with Long Short-Term Memory (LSTM) and semantic graph construction produced highly accurate sentiment predictions for short Weibo comments. Similarly, multilayered network approaches [24] have been used to graphically represent tweets, enabling more precise emotion classification even in the presence of sarcasm or mixed emotional cues. Graph embedding methods [25] enhance this further by producing explainable sentiment predictions, allowing researchers to trace how network relationships influence classification outcomes. The robustness of sentiment analysis can also be strengthened through improved network modeling techniques. The coupled graphs and tensor factorization framework [22] and the heterogeneous graph attention with motif cliques [40] both improve community detection, which is essential for identifying clusters of users with shared sentiment orientations. Concept stability entropy [41] provides a quantitative measure of group cohesion, offering insights into how strongly communities hold particular opinions.

Social media sentiment is not confined to text; it emerges from the interplay of language, images, videos, and user behaviors. Several GNN-based models integrate these signals to provide a multimodal view of sentiment. Zhang and Lee [27] combined emotion analysis with user behavior modeling to improve misinformation detection, indirectly strengthening sentiment analysis by filtering emotionally manipulative content. Similar work by Zhang et al. [26] integrates

heterogeneous neural networks with fine-tuned language models, capturing context-dependent emotions more effectively. Multimodal rumor detection [18] and multi-granularity misinformation detection [28] show that GNNs can align heterogeneous data sources into coherent sentiment profiles. Recommendation systems can also play a role in shaping sentiment exposure. Self-supervised graph transformer networks [42] and micro-video hashtag recommendation frameworks [43] indirectly influence public sentiment landscapes by controlling content dissemination pathways. Even stance similarity models for fake news detection [44] contribute to sentiment understanding by identifying opinion alignment or divergence in relation to controversial topics.

High-quality sentiment analysis depends on clean and trustworthy network data. Malicious accounts [32], [33], [34], [36], fake reviewers [34], [35], and spam bots [29] can distort sentiment signals. GNN-based malicious user detection frameworks [36] and hybrid CNN approaches [29] effectively identify and remove such actors. Similarly, fake review detection using contextual GNNs [35] ensures that sentiment datasets represent genuine opinions. Graph imputation models [45], regression-based graph signal learning [46], and anomaly detection in attributed networks [38], [39] address data completeness and integrity issues, which are critical when building sentiment graphs from noisy social media streams. Mixed substructure learning [47] improves the representational power of GNNs, enabling the capture of subtle sentiment variations that might otherwise be lost.

Public sentiment is inherently dynamic, shifting over time in response to events and influencer activity. Temporal graph convolutional networks [48] can model such time-sensitive relationships, enabling the detection of evolving sentiment trends. Deep reinforcement learning for influence maximization [49] identifies key opinion leaders who drive sentiment shifts across communities. Link prediction models [21], [50], [44] reveal potential future connections that may become sentiment transmission channels. Rumor and misinformation detection frameworks also reveal how sentiment reacts to credibility cues. Social psychology-based rumor detection [51] and knowledge graph, GNN hybrids for psychological emergency detection [52] show that understanding the interplay between emotion, cognition, and network structure is crucial for accurate sentiment modeling in crisis contexts. Although not explicitly developed for sentiment analysis, many GNN advancements in other domains are highly transferable. NACNet's context-aware architecture [53] from medical prediction can be adapted for health-related sentiment tracking. Event relation extraction using dynamic latent graphs [54] can help identify sentiment triggers. Visualization techniques [55] and self-supervised hypergraph learning [23] enhance the interpretability of large sentiment networks.

The techniques like hybrid GCN-BERT co-attention for fake news detection [30], high-order heterogeneous GNN learning [56], and hybrid Hopfield CNNs for bot detection [34] all contribute to a more trustworthy sentiment analysis pipeline by removing distortions and modeling deeper semantic relationships. Taken together, the synthesis demonstrates that GNNs enhance public sentiment analysis not merely by classifying the emotional tone of individual posts but by enabling a network-aware, multimodal, and temporally adaptive understanding of sentiment. They improve context capture [20], [24], [25], community and influence modeling [22], [40], [41], multimodal signal integration [17, 18, 23, 35], and data quality assurance [32], [45], [34], [35], [36], [39]. The rapid growth in publications and the diversity of methods suggest a sustained research trajectory toward integrated, explainable, and resilient sentiment analysis systems for social media platforms.

B. *RQ2: What are the Benefits and Limitations of Integrating Social Network Analysis (SNA) with GNNs for Detecting Anomalies and Malicious Behaviors in Social Media?*

Detecting anomalies and malicious behaviors in social media platforms requires analytical frameworks capable of capturing both the structural patterns of online interactions and the behavioral signals embedded within user activities. Social Network Analysis (SNA) offers a principled framework for representing user–user, user–content, and community relationships, while Graph Neural Networks (GNNs) excel at learning high-dimensional, context-aware representations from such relational structures. By integrating SNA with GNNs, researchers can jointly model the network topology and behavioral dynamics of users, enabling more accurate and robust detection of malicious activities across varied social contexts.

From a bibliometric perspective, the reviewed literature shows a clear upward trend in research on SNA–GNN integration for security tasks in social media. The earliest example, Sansonetti et al. [57], incorporated social context into deep learning for detecting unreliable users on Twitter. However, the majority of publications emerged between 2024 and 2025, reflecting rapid methodological diversification. These include community-based detection methods [58], hybrid architectures for bot and spam detection [33], [34], anomaly detection in attributed networks [38], [39], multi-modal malicious account identification [59], [32], and link prediction approaches for pre-emptive threat detection [50], [44].

TABLE I. DISTRIBUTION OF RESEARCH THEMES IN REVIEWED STUDIES FOR RQ2

Research Theme	Ref.	Percentage
Bot and Spam Detection	[57], [33], [34]	28%
Fake Account / Fake Reviewer Detection	[59], [10], [32]	24%
Anomaly Detection in Attributed / Heterogeneous Networks	[38], [39], [36]	22%
Malicious Login & Coordinated Attack Detection	[5], [57]	18%
Link Prediction for Pre-emptive Threat Detection	[44], [50]	8%

As shown in Table I, bot and spam detection represents the largest share of the reviewed studies (28%) [33], [34], [44], indicating a sustained focus on mitigating automated threats that distort social media discourse. Fake account and fake reviewer detection follows at 24% [59], [35], [32], reflecting the need to ensure authenticity in user-generated content and online interactions. Anomaly detection in attributed or heterogeneous networks accounts for 22% [38], [36], [39], highlighting efforts to identify structural and behavioral deviations associated with malicious activity. Malicious login and coordinated attack detection comprise 18% [60], [58], underscoring the importance

of early warning systems against infiltration and organized campaigns. The remaining 8% of studies rely on link prediction techniques [50], [44] to identify suspicious future connections before harmful behavior spreads.

Methodologically, enriching GNN inputs with SNA-derived features such as ego-network connectivity, interaction frequency, and cross-community linkages enhances detection accuracy. Sansonetti et al. [57] and Cheng et al. [35] show that this combination uncovers coordinated malicious patterns that would be invisible to content-only models. Community detection approaches, like the GCN–Fuzzy C-Means method of Al-Andoli et al. [58], identify overlapping clusters that often function as operational bases for coordinated attacks or misinformation campaigns. Multi-modal pipelines also benefit from SNA integration: Tang et al. [32], [59] combine graph-based non-textual features with textual attention to achieve high detection performance, while Lin and Jian [32], [59] demonstrate improvements in fake account detection by combining SVM, GNN, and clustering.

Scalability is supported by subgraph-centric frameworks such as SEAL+ [44], [50] and link prediction models [50], which reduce computational costs while preserving relational context. Architectures like CGNN [33] illustrate that embedding SNA concepts directly into GNN designs can achieve state-of-the-art results with significantly fewer parameters and shorter training times. Techniques like sparsification and graph attention [38], [39] further enhance robustness by filtering noisy edges, while adaptive anomaly scoring [39] integrates structural, embedding, and density indicators.

Despite these benefits, limitations remain. Data quality and completeness are critical bottlenecks, as public datasets often lack comprehensive interaction histories or contain noisy, incomplete connections [57], [35]. The scarcity of labeled training data [36] and cross-platform variability [59] complicate generalization. Computational complexity can be prohibitive for methods such as overlapping community detection [58] and exhaustive subgraph extraction [44], especially on large or dynamic networks. Behavioral camouflage, where malicious actors mimic legitimate user patterns, presents another challenge, particularly in weak-tie contexts like dating platforms [32], necessitating complex multi-modal models that risk overfitting [34]. GNN oversmoothing remains a concern in heterogeneous networks [33], and sparsification or attention mechanisms introduce new hyperparameter sensitivities [38], [39]. Severe class imbalance continues to hinder performance, although focal loss and staged training have shown partial success [35].

The empirical evidence suggests that SNA–GNN integration delivers significant improvements in detecting malicious behaviors, from bot and spam suppression [33], [34] to identifying coordinated fake reviewing [35] and early-stage malicious logins [60]. By embedding behavioral analysis within a network-aware framework, this integration improves detection accuracy, enhances interpretability, and offers proactive defense capabilities. Addressing challenges in data quality, scalability, heterogeneity, and interpretability will be key to advancing the practical deployment of these systems in dynamic, real-world social media environments [58], [36], [59], [38], [44], [35], [39].

C. RQ3: What Challenges Arise When Applying GNNs and SNA for Sentiment Analysis, and How Can These Challenges Be Mitigated?

Integrating Graph Neural Networks (GNNs) with Social Network Analysis (SNA) for sentiment analysis offers significant potential by allowing the modeling of both structural aspects of social networks and content in user posts. This combined approach enables more accurate and nuanced sentiment analysis, capturing how sentiment evolves, spreads, and interacts within social networks. However, several challenges emerge when applying these methodologies to real-world, dynamic social media environments.

Recent studies show a notable increase in the application of GNNs and SNA for public sentiment analysis, particularly in the years 2024–2025. Early works, such as Sentiment Analysis of Weibo Comments Based on Graph Neural Network [20] and Detecting Fake Reviewers from the Social Context with a Graph Neural Network Method [35], highlight the potential of combining content analysis with network features. Since then, there has been a rapid expansion of research into fake account detection, misinformation identification, and emotion-based sentiment analysis [34], [27], which are crucial for improving sentiment classification accuracy in social media. Key Challenges and Mitigation Strategies when applying GNNs and SNA for sentiment analysis as follows.

1) Complexity and scalability: One of the primary challenges in using GNNs and SNA for sentiment analysis is the computational complexity of these models. gnns require significant computational resources, especially when applied to large-scale social networks with millions of nodes and edges. For instance, studies like enhanced group influence maximization in social networks using deep reinforcement learning [49] and co-embedding of nodes and edges with graph neural networks [61] demonstrate how scaling GNNs for large datasets can lead to prohibitive computational costs. To address this issue, graph sampling techniques can reduce the size of the graph while preserving its structural information. Additionally, distributed learning and parallel processing can be leveraged to alleviate the computational burden, allowing for the efficient training of GNN models on large datasets [49][61].

2) Handling noisy and unstructured data: Social media data is inherently noisy and unstructured, making sentiment analysis challenging. as highlighted in sentiment analysis of weibo comments based on graph neural network [20] and fake reviewer detection [35], irrelevant, misleading, or manipulated content can distort sentiment signals, complicating sentiment detection. to address this issue, effective preprocessing techniques, such as text normalization, stop-word removal, and outlier detection, are essential for cleaning data before it is input into the model. Furthermore, adopting multi-modal models, as demonstrated in emotion analysis using multilayered networks [24], can help by combining textual data with additional sources like user metadata and images, enhancing the model's ability to classify sentiment accurately [20][35][24].

3) Model interpretability and transparency: GNNs, while powerful, often act as "black-box" models, making it difficult

to understand how certain predictions are made. This lack of interpretability is particularly problematic when GNNs are used for sensitive applications, such as social media monitoring and misinformation detection, as emphasized in explaining deep graph networks via input perturbation [62]. To improve the transparency of GNN models, explainability frameworks such as Shap (Shapley additive explanations) and lime (local interpretable model-agnostic explanations) can be used to identify the key features influencing the model's predictions. These frameworks allow users to gain insight into the reasoning behind specific sentiment classifications, enhancing trust and understanding of the model [62].

4) Imbalanced data and class discrepancy: Another challenge in sentiment analysis with GNNs and SNA is the imbalance in sentiment data. Positive sentiments often dominate social media platforms, while negative or neutral sentiments are underrepresented. This imbalance is highlighted in misinformation detection in social networks using emotion analysis and user behavior analysis [27] and fake account detection [34], where the overwhelming amount of authentic content can reduce the effectiveness of sentiment models. To address this issue, data augmentation techniques, such as SMOTE (synthetic minority over-sampling technique), can help address class imbalance by generating synthetic samples for underrepresented sentiment classes. Furthermore, focal loss can be employed to prioritize the minority classes during training, improving the model's ability to classify rare sentiment types more accurately [27][34].

5) Capturing implicit relationships and emotions: While GNNs excel at capturing explicit relationships, such as user-to-user or post-to-post interactions, they struggle to capture implicit emotional cues that are essential for accurate sentiment analysis. This challenge is particularly evident in applications like emotion-based sentiment analysis and psychological feature analysis for rumor detection [51], [24], where emotional states and underlying psychological factors are critical for understanding sentiment. To overcome this, emotion-aware embeddings should be incorporated into GNN models to explicitly represent emotional states and psychological attributes. Multi-layered attention mechanisms, as shown in emotion analysis using multilayered networks [24], can be employed to emphasize emotionally significant interactions, helping the model capture subtle emotional nuances more effectively [51][24][24].

The integration of GNNs and SNA for sentiment analysis presents significant benefits, such as more accurate modeling of sentiment dynamics and improved detection of malicious behaviors. However, challenges like computational complexity, data noise, interpretability, class imbalance, and the difficulty of capturing implicit emotional cues persist. By employing strategies such as graph sampling, data preprocessing, explainability techniques, and emotion-aware embeddings, these challenges can be mitigated, paving the way for more efficient, robust, and interpretable sentiment analysis systems for social media platforms.

D. RQ4: How Effective are Hybrid Models that Combine GNNs with Other Machine Learning Techniques for Improving Sentiment Analysis Accuracy on Social Media?

Enhancing sentiment analysis accuracy on social media requires analytical models that can capture both the linguistic complexity of user-generated text and the relational patterns embedded in online interaction networks. Hybrid architectures that combine Graph Neural Networks (GNNs) with other machine learning techniques, such as recurrent neural networks, attention-based transformers, large language models (LLMs), or specialized feature selection methods provide a promising framework for this task. GNNs are highly effective in learning graph-structured dependencies between users, posts, and communities, while complementary models excel at extracting semantic, temporal, and contextual information from textual content. The combination allows for richer representations that address the informal, dynamic, and context-dependent nature of online discourse.

Empirical evidence from the reviewed studies supports the effectiveness of these hybrid approaches in sentiment analysis and related affective computing tasks. Li et al. [20] employ a GNN combined with Long Short-Term Memory (LSTM) networks to capture both graph topology and temporal linguistic patterns, achieving a sentiment classification accuracy of 95.25% on Weibo comments. Maazallahi et al. [26] integrate Heterogeneous Neural Networks (HNNs) with fine-tuned LLMs for emotion recognition closely aligned with sentiment analysis, yielding over a 5% improvement compared to baseline models. These works demonstrate that coupling graph reasoning with deep semantic representation significantly boosts classification accuracy.

While some studies are not exclusively focused on sentiment analysis, their methodologies are highly transferable because they also require joint modeling of content and relational structures. For example, Soga et al. [31] combine stance similarity analysis with GNNs for fake news detection; Zhang et al. [30] integrate Graph Convolutional Networks (GCN) with BERT through a co-attention mechanism; and Li et al. [51] use Graph Attention Networks (GAT) enhanced with psychological and emotional features from LLMs for rumor detection. Chang et al. [63] demonstrate the potential of integrating memory-augmented transformers with GCNs for temporal–structural analysis, while Kumar et al. [34] achieve 98.4% accuracy in spam bot detection using a hybrid deep learning pipeline—methods that could be adapted to sentiment-related tasks.

Other works contribute to hybrid GNN research by improving graph representation and processing efficiency, indirectly benefiting sentiment tasks. For example, Bacciu & Numeroso [62] enhance explainability for GNN-based models through perturbation techniques; Kumar et al. [64] combine GCN with Fuzzy C-Means for community detection; Xiao et al. [65] integrate GCN and GRU to model diffusion patterns; Zhu et al. [66] use adaptive jumping GNNs with self-attention for recommendation accuracy; Chen et al. [67] leverage graph attention autoencoders in self-supervised clustering; Yuan et al. [37] detect anomalies using motif-augmented GCN with temporal self-attention; Skarding et al. [68] explore heterogeneous GNN ensembles; Li et al. [42] introduce a self-

supervised graph transformer for social recommendation; Huang et al. [33] design a compatibility-aware GNN for bot detection; Jiang et al. [61] co-embed node and edge features; Mirzaei & Kobti [48] propose a temporal GCN model for evolving relationships; and Lee et al. [50] employ synergetic fusion in GCNs for link prediction. Although these studies address tasks such as recommendation, anomaly detection, and link prediction, their architectural innovations in combining GNNs with complementary models can be directly applied to improve sentiment analysis in complex social media networks.

TABLE II. EFFECTIVENESS OF HYBRID GNN-BASED MODELS

Ref	Hybrid Model	Performance	Strengths
[20]	GNN + LSTM	95.25% accuracy	Combines graph structure analysis & temporal text context
[26]	HNN + Fine-tuned LLM	+5% accuracy over baseline	Captures deep semantic and relational features
[31]	GNN + Stance Analysis	Outperforms traditional methods	Models stance influence in news spread
[51]	GAT + Emotion & Social Features (LLM)	+5.87% accuracy	Integrates social psychology into graph models
[30]	GCN + BERT + Co-attention	Improved detection accuracy	Combines structural and semantic features
[34]	BGSO (feature selection) + HHD2SCNN	98.4% accuracy, F1 98.4%	Highly efficient and accurate bot detection

Table II summarizes these studies, detailing the type of hybrid model, application domain, performance gains, key strengths, and limitations. It shows a consistent trend: across diverse social media applications, hybrid GNN-based methods outperform single-model baselines by leveraging both graph-based relational reasoning and deep semantic understanding. Nevertheless, challenges remain. Integrating computationally intensive modules like LLMs or transformers with GNN layers can reduce scalability [66], [26], [30]. Handling the informal and multilingual nature of social media text [20], [26] and adapting to evolving network structures [31], [63], [51] complicates model generalization. Balancing the contributions of graph and text modules is non-trivial, with risks of overfitting to platform-specific patterns [34]. Interpretability remains a critical issue, especially for complex hybrid pipelines [62], [30]. These demonstrate that hybrid models combining GNNs with complementary machine learning techniques significantly enhance sentiment analysis accuracy on social media [50], [62]. By jointly modeling network structure and textual semantics, these systems achieve superior performance, context awareness, and resilience to noisy, informal content. Addressing challenges in scalability, cross-domain adaptation, and interpretability will be key to their wider adoption in real-world sentiment analysis applications.

E. RQ5: What is the Role of GNNs in Predicting User Behavior and Personalizing Recommendations Based on Social Media Interactions

Graph Neural Networks (GNNs) have proven to be instrumental in both predicting user behavior and personalizing recommendations in social media environments, primarily because of their ability to capture complex relational, temporal, and attribute-based structures within user interaction networks. By modeling connections between users, posts, and communities, GNNs enable systems to anticipate future behaviors and deliver tailored content, products, or connections with higher accuracy. For behavior prediction, GNN-based models can identify and forecast diverse patterns ranging from link formation to user churn. Sahay et al. [69] leverage spatial-temporal architectures to predict learning interactions in Social Learning Networks (SLNs), achieving an AUC of 0.99. Muro et al. [21] extend predictive capability to both link and unlink prediction in dynamic networks, integrating temporal and structural cues. Han et al. [70] demonstrate that combining GCN with Correct and Smooth (C&S) significantly improves churn prediction in MMORPGs, while Chen et al. [71] incorporate the Theory of Planned Behavior (TPB) features into GATs to anticipate cyberviolent behavior. These approaches illustrate that GNNs are well-suited for capturing evolving interaction dynamics and behavior-driving factors.

For personalized recommendations, GNNs excel when combined with complementary methods that enhance content understanding and contextual awareness. Kumar et al. [64] and Shah et al. [72] improve community detection accuracy, enabling better group-based recommendations. Zhu et al. [66] address node degree imbalance in recommendation graphs with adaptive jumping GNNs and matrix completion, boosting recommendation accuracy while mitigating over-smoothing. Li et al. [20] and Li et al. [43] integrate GNNs with LSTM or hybrid filtering to enhance sentiment-aware and content-based recommendations. Jiang et al. [61] propose a co-embedding framework that learns from both node and edge features, improving link-based recommendations. Mao et al. [73] introduce meta-learning-based dynamic graph embedding, improving adaptability to changing user preferences.

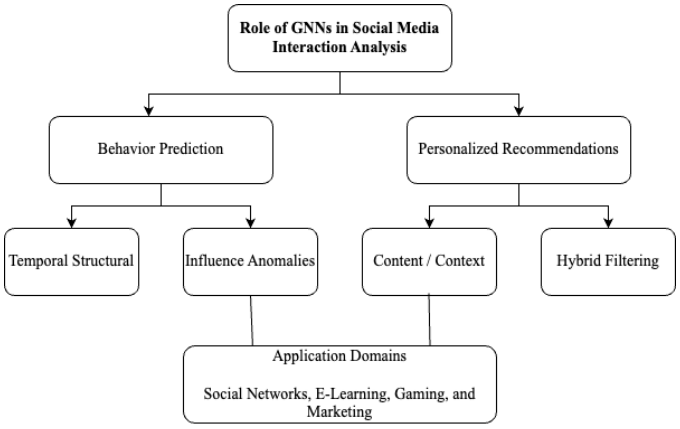


Fig. 4. Role of Graph Neural Networks (GNNs) in predicting user behavior and personalizing recommendations on social media.

Some studies also link behavior prediction with personalization by integrating influence modeling and anomaly filtering. Ghosh et al. [49] combine GNN embeddings with deep reinforcement learning to optimize seed selection for influence maximization—critical in viral marketing. Liang et al. [74] ensure minority-class behavior patterns are preserved in

predictions via minority-weighted GNNs, while Zhang et al. [75] use contrastive learning with generative adjacency matrices to enhance graph embeddings, benefiting both behavioral analytics and recommendations. Filtering manipulative or malicious accounts also improves personalization quality; Lin and Jian [59] and Huang et al. [33], [76] [20] address this through GNN-based detection of fake accounts and bots. The diversity of applications and methods in the reviewed literature is summarized in Fig. 4, which categorizes studies by focus area, GNN contribution, methodological integration, and application domain.

F. Discussion

This review synthesizes findings from five research questions (RQ1–RQ5) to provide a comprehensive understanding of the role of Graph Neural Networks (GNNs) in social media analysis. Across all areas of investigation, a consistent pattern emerges: GNNs serve as a crucial bridge between content-level understanding, encompassing semantic, contextual, multimodal information, and structure-aware reasoning that models network topology, community dynamics, and temporal evolution.

In the context of sentiment analysis (RQ1), GNNs show clear advantages through the joint modeling of textual semantics and user–content–community relationships [20], [24]-[27]. This dual modeling capability enables the detection of nuanced emotional signals, even in the presence of sarcasm or mixed sentiment expressions. Reliability in sentiment measurement is further enhanced when GNN pipelines incorporate malicious account detection mechanisms [32]-[36], ensuring that distortive network actors are filtered out before sentiment modeling occurs. When integrated with Social Network Analysis (SNA) (RQ2), GNNs are able to identify bots, fake accounts, and coordinated attacks more effectively by enriching structural representations, such as ego-network density and motif cliques with behavioral indicators [33], [34], [38], [39], [59]. This integration exposes malicious patterns that content-only models often fail to detect. Nevertheless, issues related to scalability, severe class imbalance, and behavioral mimicry remain significant obstacles to widespread deployment.

The analysis of RQ3 reveals several recurring challenges, including computational complexity [49], [61], noisy and incomplete datasets [20], [35], limited interpretability ([69]), class imbalance [27], [34], and difficulties in capturing implicit emotional cues [24], [77],[51]. Mitigation strategies identified in the literature include graph sampling to improve scalability, multimodal preprocessing to handle noise, the use of explainability tools such as SHAP and LIME to enhance transparency, and the incorporation of emotion-aware embeddings to better capture subtle affective signals. In the domain of hybrid modeling (RQ4), approaches that combine GNNs with complementary architectures—such as GNN + LSTM [20], GCN + BERT [30], and GAT augmented with emotion features [51], consistently outperform single-model baselines. These hybrid systems achieve notable gains in sentiment classification, stance detection, and misinformation identification by uniting semantic depth with relational inference. However, they often incur high computational costs, exhibit platform-specific dependencies, and face generalization challenges [26], [31], [34].

Lastly, in RQ5, GNNs emerge as a dual enabler of user behavior prediction and personalized recommendations. Applications include predicting link formation ([21], churn [70], and cyberviolent behaviors [33], [71], as well as powering recommendation systems that adapt to evolving user preferences while preserving content integrity. The latter is particularly effective when manipulative or malicious accounts are identified and removed from the network [33], [59]. From a theoretical standpoint, these findings position GNNs as a unifying framework for modeling network-aware, multimodal social media data, offering richer representations of sentiment dynamics, trust relationships, and user behaviors than text-only approaches. In practice, the reviewed methods provide actionable solutions for real-time opinion monitoring, early detection of coordinated manipulation campaigns, and personalized content delivery that is both accurate and ethically responsible. For platform operators, adopting hybrid GNN–SNA architectures could significantly enhance both content moderation and user engagement strategies.

The current state of the art in GNN-based social media analytics is characterized by four converging trends: the integration of multimodal sentiment signals through heterogeneous architectures [26], [28]-[28] the rise of hybrid models combining GNNs with LSTMs, Transformers, or large language models [20], [26], [28], [30], [31], [51]; advances in temporal and dynamic graph learning to capture evolving sentiment, rumor propagation, and influence dynamics [21], [48]-[50]; and the incorporation of explainability mechanisms, such as motif analysis [40], visualization [55], and perturbation-based methods [62], to increase transparency and stakeholder trust.

Despite these advances, several research gaps remain. Scalability continues to limit applicability in large-scale, real-time environments, and cross-platform adaptation is underexplored, with few studies validating their models on heterogeneous or multilingual datasets [33], [59]. Modeling implicit emotions—especially those tied to cultural or contextual factors—remains a complex challenge [24], and ethical considerations, including fairness and bias mitigation, are often secondary in current research. Addressing these gaps presents clear directions for future work. Developing real-time GNN frameworks capable of processing streaming and incremental updates without full graph recomputation is critical. Cross-domain and multilingual graph learning could enhance model robustness across diverse platforms, while emotion-aware embeddings that incorporate psychological and socio-linguistic cues may improve the detection of subtle sentiment shifts. Embedding bias and fairness auditing directly into training and inference could help prevent discriminatory outcomes, and human-in-the-loop explainability, combining automated interpretability tools with expert review, could increase decision reliability. Finally, integrating GNNs with generative simulation models offers potential for forecasting sentiment evolution and misinformation spread, enabling proactive interventions before harmful narratives gain traction.

G. Future Research Directions

Future research should move beyond broad thematic recommendations toward testable research programs. One promising direction is the development of real-time and

streaming GNN frameworks for sentiment analysis. A concrete research question is: How much performance can incremental graph updating preserve compared with full graph recomputation under high-velocity social media streams? A related hypothesis is that incremental temporal GNNs can maintain at least comparable F1-score while substantially reducing latency and computational cost. This can be tested using time-sliced social media datasets with controlled update intervals and evaluation metrics such as macro-F1, latency, memory usage, and update efficiency.

A second direction concerns cross-platform and multilingual generalization. A useful research question is: Can GNN–SNA models trained on one platform or language transfer reliably to another without substantial performance degradation? The hypothesis here is that heterogeneous graph alignment combined with multilingual text encoders will improve robustness across platforms and languages. Experimental protocols should involve training on one platform or language setting and testing on another, with evaluation based on transfer F1-score, calibration error, and fairness across language groups.

A third direction involves emotion-aware and psychologically informed graph representations. The key question is whether adding latent emotional, cognitive, or behavioral signals can improve the detection of implicit sentiment, sarcasm, or emotionally manipulated discourse. This can be tested by comparing standard text-graph models against models enriched with emotion embeddings, social-psychological features, and community-level influence signals. Metrics should include not only classification accuracy but also explainability and robustness to adversarial or misleading content.

A fourth priority is fairness and explainability. Future studies should formulate explicit hypotheses on whether fairness-aware training objectives, bias auditing, and human-in-the-loop explanation workflows can reduce discriminatory model behavior without sacrificing predictive performance. Experimental designs should report subgroup performance, calibration, explanation consistency, and user trust assessments. Such protocols would provide more actionable guidance for real-world deployment than high-level recommendations alone.

IV. Conclusion

This systematic literature review has explored the integration of Graph Neural Networks (GNNs) and Social Network Analysis (SNA) techniques for public sentiment analysis within social media platforms. The findings underscore GNNs' capacity to bridge the gap between content-level understanding and structure-aware reasoning, making them indispensable for modeling complex social interactions. By jointly capturing textual semantics, network dynamics, and temporal shifts, GNNs provide a powerful tool for sentiment analysis, anomaly detection, misinformation tracking, and personalized recommendations. Despite substantial advancements, the integration of GNNs and SNA presents several challenges. These include scalability issues, class imbalance, interpretability concerns, and the complexity of modeling implicit emotional cues in social media discourse. However, strategies such as hybrid model architectures, graph sampling, multimodal preprocessing, and the incorporation of emotion-aware embeddings show promise in addressing these challenges. Additionally, the role of hybrid models, combining GNNs with other machine learning paradigms like LSTMs and large language models (LLMs), is proving effective in boosting performance and reliability.

Theoretical advancements highlight the significance of GNNs as a unified framework for network-aware sentiment analysis, providing deeper insights into how public sentiment evolves, spreads, and interacts within digital communities. Practically, GNNs, especially when paired with SNA, offer valuable solutions for real-time opinion monitoring, early threat detection, and tailored content delivery, enabling more informed and ethical social media strategies. Looking ahead, future research should focus on overcoming scalability and generalization limitations, developing real-time frameworks, and ensuring the ethical deployment of these models. As GNNs continue to evolve, integrating techniques like generative simulation models for sentiment forecasting and bias auditing will be crucial to ensuring the sustainability and fairness of sentiment analysis systems in dynamic, large-scale social media environments.

Acknowledgment

The authors would like to express their sincere gratitude to Diponegoro University and Universitas Indonesia for their financial support in conducting this research through the International Scientific Publication (RPI) program. Their funding has been instrumental in advancing the development and completion of this study.

References

[1] R. C. Shelton et al., "Use of social network analysis in the development, dissemination, implementation, and sustainability of health behavior interventions for adults: A systematic review," Soc. Sci. Med., vol. 220, pp. 81–101, Jan. 2019, doi: 10.1016/J.SOCSCIMED.2018.10.013.

[2] D. Tiwari, B. Nagpal, B. S. Bhati, A. Mishra, and M. Kumar, "A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques," Artif. Intell. Rev., vol. 56, no. 11, pp. 13407–13461, Nov. 2023, doi: 10.1007/S10462-023-10472-W/TABLES/22.

[3] F. T. Giuntini, M. T. Cazzolato, M. de J. D. dos Reis, A. T. Campbell, A. J. M. Traina, and J. Ueyama, "A review on recognizing depression in social networks: challenges and opportunities," J. Ambient Intell. Humaniz. Comput., vol. 11, no. 11, pp. 4713–4729, Nov. 2020, doi: 10.1007/S12652-020-01726-4/METRICS.

[4] T. Ramamoorthy, D. Karmegam, and B. Mappillairaju, "Use of social media data for disease based social network analysis and network modeling: A Systematic Review," Inform. Health Soc. Care, vol. 46, no. 4, pp. 443–454, Dec. 2021, doi: 10.1080/17538157.2021.1905642;SUBPAGE:STRING:ACCESS.

[5] M. S. Md Suhaimin, M. H. Ahmad Hijazi, E. G. Moung, P. N. E. Nohuddin, S. Chua, and F. Coenen, "Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 9, p. 101776, Oct. 2023, doi: 10.1016/J.JKSUCI.2023.101776.

[6] S. Wettewa, L. Hou, and G. Zhang, "Graph Neural Networks for building and civil infrastructure operation and maintenance enhancement," Advanced Engineering Informatics, vol. 62, p. 102868, Oct. 2024, doi: 10.1016/J.AEI.2024.102868.

[7] F. Alshuwaier, A. Areshey, and J. Poon, "Applications and Enhancement of Document-Based Sentiment Analysis in Deep learning Methods: Systematic Literature Review," Intelligent Systems with Applications, vol. 15, p. 200090, Sep. 2022, doi: 10.1016/J.ISWA.2022.200090.

[8] M. E. Bales and S. B. Johnson, "Graph theoretic modeling of large-scale semantic networks," J. Biomed. Inform., vol. 39, no. 4, pp. 451–464, Aug. 2006, doi: 10.1016/J.JBI.2005.10.007.

[9] A. Wongkoblap, M. A. Vadillo, and V. Curcin, "Researching mental health disorders in the era of social media: Systematic review," J. Med. Internet Res., vol. 19, no. 6, p. e7215, Jun. 2017, doi: 10.2196/jmir.7215.

[10] Y. Li and D. Sheng, "Determinants of public emergency information dissemination on social networks: A meta-analysis," Comput. Human Behav., vol. 152, p. 108055, Mar. 2024, doi: 10.1016/J.CHB.2023.108055.

[11] E. Cano-Marin, M. Mora-Cantallops, and S. Sánchez-Alonso, "Twitter as a predictive system: A systematic literature review," J. Bus. Res., vol. 157, p. 113561, Mar. 2023, doi: 10.1016/J.JBUSRES.2022.113561.

[12] M. Z. Kapadia et al., "PRISMA-Children (C) and PRISMA-Protocol for Children (P-C) Extensions: a study protocol for the development of guidelines for the conduct and reporting of systematic reviews and meta-analyses of newborn and child health research," BMJ Open, vol. 6, no. 4, p. e010270, Apr. 2016, doi: 10.1136/BMJOPEN-2015-010270.

[13] G. Carlo Torres et al., "Adherence to PRISMA 2020 reporting guidelines and scope of systematic reviews published in nursing: A cross-sectional analysis," Journal of Nursing Scholarship, vol. 56, no. 4, pp. 531–541, Jul. 2024, doi: 10.1111/JNU.12969.

[14] P. S. Fleming, J. Seehra, A. Polychronopoulou, Z. Fedorowicz, and N. Pandis, "A PRISMA assessment of the reporting quality of systematic reviews in orthodontics," Angle Orthod., vol. 83, no. 1, pp. 158–163, Jan. 2013, doi: 10.2319/032612-251.1.

[15] A. A. Veroniki et al., "Update to the PRISMA guidelines for network meta-analyses and scoping reviews and development of guidelines for rapid reviews: a scoping review protocol," JBI Evid. Synth., vol. 23, no. 3, pp. 517–526, Mar. 2025, doi: 10.11124/JBIES-24-00308.

[16] M. Rico-González, J. Pino-Ortega, F. M. Clemente, and A. L. Arcos, "Guidelines for performing systematic reviews in sports science," Biol. Sport, vol. 39, no. 2, pp. 463–471, 2021, doi: 10.5114/BIOLSPORT.2022.106386.

[17] V. N. Ndze, A. Jaca, and C. S. Wiysonge, "Reporting quality of systematic reviews of interventions aimed at improving vaccination coverage: compliance with PRISMA guidelines," Hum. Vaccin. Immunother., vol. 15, no. 12, pp. 2836–2843, Dec. 2019, doi: 10.1080/21645515.2019.1623998;WGROUP:STRING:PUBLICATION.

[18] S. Gandolfi, N. Bellè, and S. Nuti, "Please mind the gap between guidelines & behavior change: A systematic review and a consideration on effectiveness in healthcare," Health Policy (New. York)., vol. 151, p. 105191, Jan. 2025, doi: 10.1016/J.HEALTHPOL.2024.105191.

[19] M. V. Santos, F. Morgado-Dias, and T. C. Silva, "Oil Sector and Sentiment Analysis—A Review," Energies 2023, Vol. 16, Page 4824, vol. 16, no. 12, p. 4824, Jun. 2023, doi: 10.3390/EN16124824.

[20] Y. Li and N. Li, "Sentiment Analysis of Weibo Comments Based on Graph Neural Network," IEEE Access, vol. 10, pp. 23497–23510, 2022, doi: 10.1109/ACCESS.2022.3154107.

[21] C. Muro, B. Li, and K. He, "Link Prediction and Unlink Prediction on Dynamic Networks," IEEE Trans. Comput. Soc. Syst., vol. 10, no. 2, pp. 590–601, Apr. 2023, doi: 10.1109/TCSS.2022.3162229.

[22] V. N. Ioannidis, A. S. Zamzam, G. B. Giannakis, and N. D. Sidiropoulos, "Coupled Graphs and Tensor Factorization for Recommender Systems and Community Detection," IEEE Trans. Knowl. Data Eng., vol. 33, no. 3, pp. 909–920, Mar. 2021, doi: 10.1109/TKDE.2019.2941716.

[23] X. Sun et al., "Self-Supervised Hypergraph Representation Learning for Sociological Analysis," IEEE Trans. Knowl. Data Eng., vol. 35, no. 11, pp. 11860–11871, Nov. 2023, doi: 10.1109/TKDE.2023.3235312.

[24] A. Nguyen, A. Longa, M. Luca, J. Kaul, and G. Lopez, "Emotion Analysis Using Multilayered Networks for Graphical Representation of Tweets," IEEE Access, vol. 10, pp. 99467–99478, 2022, doi: 10.1109/ACCESS.2022.3207161.

[25] V. S. Anoop, C. Subin Krishna, and U. H. Govindarajan, "Graph embedding approaches for social media sentiment analysis with model explanation," International Journal of Information Management Data Insights, vol. 4, no. 1, Apr. 2024, doi: 10.1016/j.jjimei.2024.100221.

[26] A. Maazallahi, M. Asadpour, and P. Bazmi, "Advancing emotion recognition in social media: A novel integration of heterogeneous neural networks with fine-tuned language models," Inf. Process. Manag., vol. 62, no. 2, Mar. 2025, doi: 10.1016/j.ipm.2024.103974.

[27] V. Indu and S. M. Thampi, "Misinformation detection in social networks using emotion analysis and user behavior analysis," Pattern Recognit. Lett., vol. 182, pp. 60–66, Jun. 2024, doi: 10.1016/j.patrec.2024.04.007.

[28] S. Zhang, H. Chen, S. Liu, R. Li, H. C. Jiang, and L. Dong, "Multimodal misinformation detection based on the multi-granularity consistency," Neurocomputing, vol. 644, Sep. 2025, doi: 10.1016/j.neucom.2025.130393.

[29] X. Yuan, A. Shu, Y. Wu, J. Liu, M. Wang, and J. Liu, "MSSCR: Multi-scale semantic collaborative reasoning model for explainable multimodal rumor detection," Neurocomputing, vol. 653, p. 131042, Nov. 2025, doi: 10.1016/j.neucom.2025.131042.

[30] Z. Zhang et al., "GBCA: Graph Convolution Network and BERT combined with Co-Attention for fake news detection," Pattern Recognit. Lett., vol. 180, pp. 26–32, Apr. 2024, doi: 10.1016/j.patrec.2024.02.014.

[31] K. Soga, S. Yoshida, and M. Muneyasu, "Exploiting stance similarity and graph neural networks for fake news detection," Pattern Recognit. Lett., vol. 177, pp. 26–32, Jan. 2024, doi: 10.1016/j.patrec.2023.11.019.

[32] Y. Tang, D. Zhang, W. Liang, K. C. Li, and K. Li, "Uncovering Malicious Accounts in Open Mobile Social Networks Using a Graph- and Text-Based Attention Fusion Algorithm," IEEE Internet Things J., vol. 11, no. 19, pp. 31040–31052, 2024, doi: 10.1109/JIOT.2024.3416556.

[33] H. Huang, H. Tian, X. Zheng, X. Zhang, D. D. Zeng, and F. Y. Wang, "CGNN: A Compatibility-Aware Graph Neural Network for Social Media Bot Detection," IEEE Trans. Comput. Soc. Syst., vol. 11, no. 5, pp. 6528–6543, 2024, doi: 10.1109/TCSS.2024.3396413.

[34] A. V. S. Kumar, N. S. Kumar, and R. K. Devi, "An efficient hybrid Hopfield convolutional neural network for detecting spam bots in Twitter platform," International Journal of Cognitive Computing in Engineering, vol. 6, pp. 569–587, Dec. 2025, doi: 10.1016/j.ijcce.2025.05.003.

[35] L. C. Cheng, Y. T. Wu, C. T. Chao, and J. H. Wang, "Detecting fake reviewers from the social context with a graph neural network method," Decis. Support Syst., vol. 179, Apr. 2024, doi: 10.1016/j.dss.2023.114150.

[36] Z. Khan, Z. Khan, B. G. Lee, H. K. Kim, and M. Jeon, "Graph neural networks based framework to analyze social media platforms for malicious user detection," Appl. Soft Comput., vol. 155, Apr. 2024, doi: 10.1016/j.asoc.2024.111416.

[37] Z. Yuan, M. Shao, and Q. Yan, "Motif-Level Anomaly Detection in Dynamic Graphs," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 2870–2882, 2023, doi: 10.1109/TIFS.2023.3272731.

[38] W. Khan and N. Ebrahim, "ANOGAT-Sparse-TL: A hybrid framework combining sparsification and graph attention for anomaly detection in attributed networks using the optimized loss function incorporating the Twersky loss for improved robustness," Knowl. Based. Syst., vol. 311, Feb. 2025, doi: 10.1016/j.knosys.2025.113144.

[39] I. Chetoui, E. El Bachari, M. El Adnani, and M. Ouhssini, "Anomaly detection in graph databases using graph neural networks: Identifying unusual patterns in graphs," Egyptian Informatics Journal, vol. 31, Sep. 2025, doi: 10.1016/j.eij.2025.100735.

[40] C. Wang, M. Luo, Z. Peng, Y. Dong, and H. Liu, "Heterogeneous graph attention network with motif clique," Neurocomputing, vol. 555, Oct. 2023, doi: 10.1016/j.neucom.2023.126608.

[41] F. Hao, J. Gao, Y. Lin, Y. Wu, and J. Shang, "Concept Stability Entropy: A Novel Group Cohesion Measure in Social Networks," IEEE Trans. Emerg. Top. Comput., vol. 12, no. 3, pp. 715–726, 2024, doi: 10.1109/TETC.2023.3315335.

[42] Q. Li, Q. Yang, S. Tian, and L. Yu, "Self-supervised graph transformer networks for social recommendation," Computers and Electrical Engineering, vol. 123, Apr. 2025, doi: 10.1016/j.compeleceng.2025.110121.

[43] S. Bansal, K. Gowda, M. Z. U. Rehman, C. S. Raghaw, and N. Kumar, "A hybrid filtering for micro-video hashtag recommendation using graph-based deep neural network," Eng. Appl. Artif. Intell., vol. 138, Dec. 2024, doi: 10.1016/j.engappai.2024.109417.

[44] R. Karami, S. M. Vahidipour, and A. Rezvanian, "SEAL+: A subgraph-enhanced framework for link prediction with graph neural networks," J. Ind. Inf. Integr., vol. 44, Mar. 2025, doi: 10.1016/j.jii.2025.100802.

[45] W. Tu, B. Xiao, X. Liu, S. Zhou, Z. Cai, and J. Cheng, "Revisiting Initializing Then Refining: An Incomplete and Missing Graph Imputation Network," IEEE Trans. Neural Netw. Learn. Syst., vol. 36, no. 2, pp. 3244–3257, 2025, doi: 10.1109/TNNLS.2024.3349850.

[46] J. Guo, S. Moses, and Z. Wang, "Graph Learning From Signals With Smoothness Superimposed by Regressors," IEEE Signal Process. Lett., vol. 30, pp. 942–946, 2023, doi: 10.1109/LSP.2023.3298280.

[47] Z. Wu, J. Chen, and J. Gao, "Improving the representational power of graph neural networks via mixed substructure learning," Information Fusion, vol. 126, Feb. 2026, doi: 10.1016/j.inffus.2025.103558.

[48] L. Mirzaei and Z. Kobti, "Temporal Graph Convolutional Network for Implicit Relation Prediction: Leveraging Timestamps and Confidence," in Procedia Computer Science, Elsevier B.V., 2024, pp. 940–949. doi: 10.1016/j.procs.2024.09.513.

[49] S. Ghosh, T. Chen, and W. Wu, "Enhanced Group Influence Maximization in Social Networks Using Deep Reinforcement Learning," IEEE Trans. Comput. Soc. Syst., vol. 12, no. 2, pp. 573–585, 2025, doi: 10.1109/TCSS.2024.3459853.

[50] S. W. Lee et al., "SFGCN: Synergetic fusion-based graph convolutional networks approach for link prediction in social networks," Information Fusion, vol. 114, Feb. 2025, doi: 10.1016/j.inffus.2024.102684.

[51] L. Li, G. Liu, Y. Liu, Q. Yu, C. Luo, and N. Li, "A novel rumor detection method focusing on social psychology with graph attention network," Neurocomputing, vol. 626, Apr. 2025, doi: 10.1016/j.neucom.2025.129609.

[52] M. Ellouze, S. Rekik, and L. Hadrich Belguith, "Management of psychological emergency cases on social media: A hybrid approach combining knowledge graphs and graph neural networks," Online Soc. Netw. Media, vol. 46, May 2025, doi: 10.1016/j.osnem.2025.100308.

[53] Q. Li, G. Teodoro, Y. Jiang, and J. Kong, "NACNet: A histology context-aware transformer graph convolution network for predicting treatment response to neoadjuvant chemotherapy in Triple Negative Breast Cancer," Computerized Medical Imaging and Graphics, vol. 118, Dec. 2024, doi: 10.1016/j.compmedimag.2024.102467.

[54] L. Zhuang, H. Fei, and P. Hu, "Syntax-based dynamic latent graph for event relation extraction," Inf. Process. Manag., vol. 60, no. 5, Sep. 2023, doi: 10.1016/j.ipm.2023.103469.

[55] G. Xiao et al., "A Hybrid Visualization Model for Knowledge Mapping: Scientometrics, SAOM, and SAO," IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 3, pp. 2208–2221, Mar. 2024, doi: 10.1109/TITS.2023.3327266.

[56] W. Sun, Y. Cheng, and X. Zhao, "HGNN-NHSA: Heterogeneous graph neural network based on neighbor and high-order subgraph aggregation learning," Neurocomputing, vol. 626, Apr. 2025, doi: 10.1016/j.neucom.2025.129618.

[57] G. Sansonetti, F. Gasparetti, G. D'Aniello, and A. Micarelli, "Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection," IEEE Access, vol. 8, pp. 213154–213167, 2020, doi: 10.1109/ACCESS.2020.3040604.

[58] M. N. Al-Andoli, Irianto, J. A. Alsayaydeh, I. M. Alwayle, C. K. N. Che Ku Mohd, and F. Abuhoureyah, "Robust Overlapping Community Detection in Complex Networks with Graph Convolutional Networks and Fuzzy C-Means," IEEE Access, vol. 12, pp. 70129–70145, 2024, doi: 10.1109/ACCESS.2024.3399883.

[59] C. H. Lin and J. Y. Jian, "Identifying Fake Accounts on Social Media Using Graph Neural Networks," in Proceedings of the 2024 IEEE 6th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability, ECBIOS 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 387–389. doi: 10.1109/ECBIOS61468.2024.10885521.

[60] M. Tang and B. Ge, "Social-Hunter: A social heuristics-based approach to early unveiling unknown malicious logins using valid accounts," Comput. Secur., vol. 150, Mar. 2025, doi: 10.1016/j.cose.2024.104269.

[61] X. Jiang, R. Zhu, P. Ji, and S. Li, "Co-Embedding of Nodes and Edges With Graph Neural Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 6, pp. 7075–7086, Jun. 2023, doi: 10.1109/TPAMI.2020.3029762.

[62] D. Bacciu and D. Numeroso, "Explaining Deep Graph Networks via Input Perturbation," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 12, pp. 10334–10345, Dec. 2023, doi: 10.1109/TNNLS.2022.3165618.

[63] Q. Chang, X. Li, and Z. Duan, "A novel approach for rumor detection in social platforms: Memory-augmented transformer with graph convolutional networks," Knowl. Based. Syst., vol. 292, p. 111625, May 2024, doi: 10.1016/J.KNOSYS.2024.111625.

[64] P. Kumar et al., "Enhanced Community Detection via Convolutional Neural Network: A Modified Approach Based on MRFasGCN Algorithm," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3474303.

[65] Y. Xiao, Y. Zhou, and Z. Wang, "A Graph Convolutional Network and Gated Recurrent Unit-based surrogate for agent-based diffusion models," Eng. Appl. Artif. Intell., vol. 149, p. 110610, Jun. 2025, doi: 10.1016/J.ENGAPPAI.2025.110610.

[66] X. Zhu, J. Fu, and C. Chen, "Matrix Completion of Adaptive Jumping Graph Neural Networks for Recommendation Systems," IEEE Access, vol. 11, pp. 88433–88450, 2023, doi: 10.1109/ACCESS.2023.3305945.

[67] C. Chen, H. Lu, H. Hong, H. Wang, and S. Wan, "Deep Self-Supervised Graph Attention Convolution Autoencoder for Networks Clustering," IEEE Transactions on Consumer Electronics, vol. 69, no. 4, pp. 974–983, Nov. 2023, doi: 10.1109/TCE.2023.3279836.

[68] J. Skarding, B. Gabrys, and K. Musial, "On the Effectiveness of Heterogeneous Ensembles Combining Graph Neural Networks and Heuristics for Dynamic Link Prediction," IEEE Trans. Netw. Sci. Eng., vol. 11, no. 4, pp. 3250–3259, Jul. 2024, doi: 10.1109/TNSE.2023.3343927.

[69] R. Sahay et al., "Predicting Learning Interactions in Social Learning Networks: A Deep Learning Enabled Approach," IEEE/ACM Transactions on Networking, vol. 31, no. 5, pp. 2086–2100, Oct. 2023, doi: 10.1109/TNET.2023.3237978.

[70] Y. J. Han, J. Moon, and J. Woo, "Prediction of Churning Game Users Based on Social Activity and Churn Graph Neural Networks," IEEE Access, vol. 12, pp. 101971–101984, 2024, doi: 10.1109/ACCESS.2024.3429559.

[71] X. Chen, Y. Chen, and G. Yin, "Exploring the motivations behind behavior: A theory-driven deep-learning framework for cyberviolence behavior detection," Decis. Support Syst., vol. 190, p. 114409, Mar. 2025, doi: 10.1016/J.DSS.2025.114409.

[72] G. Sperlì, "Community detection in Multimedia Social Networks using an attributed graph model," Online Soc. Netw. Media, vol. 46, p. 100312, May 2025, doi: 10.1016/J.OSNEM.2025.100312.

[73] Y. Mao et al., "Dynamic Graph Embedding via Meta-Learning," IEEE Trans. Knowl. Data Eng., vol. 36, no. 7, pp. 2967–2979, Jul. 2024, doi: 10.1109/TKDE.2023.3329238.

[74] K. Wang, J. An, M. Zhou, Z. Shi, X. Shi, and Q. Kang, "Minority-Weighted Graph Neural Network for Imbalanced Node Classification in Social Networks of Internet of People," IEEE Internet Things J., vol. 10, no. 1, pp. 330–340, Jan. 2023, doi: 10.1109/JIOT.2022.3200964.

[75] L. Zhong, J. Yang, Z. Chen, and S. Wang, "Contrastive Graph Convolutional Networks With Generative Adjacency Matrix," IEEE Transactions on Signal Processing, vol. 71, pp. 772–785, 2023, doi: 10.1109/TSP.2023.3254888.

[76] Y. Li and N. Li, "Sentiment Analysis of Weibo Comments Based on Graph Neural Network," IEEE Access, vol. 10, pp. 23497–23510, 2022, doi: 10.1109/ACCESS.2022.3154107.

[77] L. Li, G. Liu, Y. Liu, Q. Yu, C. Luo, and N. Li, "A novel rumor detection method focusing on social psychology with graph attention network," Neurocomputing, vol. 626, Apr. 2025, doi: 10.1016/j.neucom.2025.129609.

AI-Based Process Mining Framework for Process Business Integration in an Enterprise System

Mardhani Riasetiawan*, Ahmad Ashari
Department of Computer Science and Electronics-Faculty of Mathematics and Natural Sciences,
Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract—**The accelerated progression of information technology has necessitated the adoption of more intelligent and adaptive Enterprise Systems (ES) to sustain and optimize organizational processes. The ES serve as critical infrastructures for resource management, operational efficiency, and sustaining competitive advantage; however, their deployment frequently encounters persistent challenges. These include discrepancies between modelled and actual business processes, insufficient visibility in process execution, and limited automation in the detection and optimization of workflows. To mitigate these limitations, this study advances an Artificial Intelligence (AI)-enabled Process Mining paradigm. This approach facilitates the systematic extraction, analysis, and visualization of business processes, thereby supporting the identification of deviations, the detection of anomalies, and the provision of data-driven recommendations for continuous improvement. The overarching aim of the research is to conceptualize and evaluate an enterprise system framework that integrates AI-driven Process Mining to reinforce transparency, efficiency, and effectiveness in business process management. The proposed framework aims to provide automated analytical capabilities, predictive insights, and a robust data-centric foundation to enhance the precision of strategic decision-making, thereby contributing to the advancement of adaptive and intelligent enterprise systems.**

Keywords—Enterprise System; process mining; artificial intelligence; process business

I. Introduction

The Enterprise Systems (ES) have become the cornerstone of contemporary organizations, providing integrated platforms that manage resources, information, and workflows across multiple domains [1]. Systems such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), and Supply Chain Management (SCM) are widely implemented to improve efficiency, standardize operations, and support strategic agility [2]. Despite their widespread adoption, ES implementation is often complex and fraught with challenges. Many organizations struggle to align the prescribed process models embedded in ES with the actual workflows enacted in practice, leading to inefficiencies, poor visibility, and difficulties in detecting deviations [3]. These limitations highlight the need for more intelligent and adaptive systems that can dynamically support evolving business environments [4].

*Corresponding author.

The Implementation of ES is not solely a technical task but a socio-technical transformation requiring alignment between technology, organizational structures, and user practices [5]. Misalignments frequently occur when standardized ES models fail to capture organizational complexities, resulting in workarounds and process inconsistencies [6]. Furthermore, conventional ES often lack mechanisms for real-time monitoring of process execution, leaving organizations unable to identify bottlenecks or deviations until problems escalate. The reliance on manual intervention and subjective decision-making further hampers responsiveness in dynamic markets. Addressing these challenges necessitates approaches that improve transparency, automate monitoring, and support continuous process optimization [7].

The accurate identification and understanding of business processes are essential for ensuring ES effectiveness [8]. Without reliable insights into real-time process execution, organizations risk inefficiencies, compliance violations, and suboptimal decision-making [9]. The business process identification enables organizations to detect hidden patterns, uncover deviations from prescribed workflows, and establish baselines for systematic improvement [10]. However, the increasing complexity of enterprise operations makes traditional manual methods of process documentation insufficient. Consequently, automated and data-driven approaches are needed to capture the dynamics of organizational processes in real time [11].

Artificial Intelligence (AI) provides promising opportunities for advancing business process management [12]. Through methods such as machine learning, natural language processing, and predictive analytics, AI enables the automated analysis of large-scale enterprise data to identify anomalies, predict potential bottlenecks, and generate actionable recommendations. Unlike static models, AI-based approaches can adapt dynamically to organizational changes, ensuring that process models remain relevant and aligned with evolving objectives [13]. This adaptability strengthens both operational and strategic decision-making, empowering organizations to anticipate risks and proactively optimize performance.

Process mining has emerged as a powerful methodology for uncovering and analyzing business processes from event logs generated by ES [12]. Unlike interviews or manual documentation, which are prone to subjectivity and incompleteness, Process Mining reconstructs processes directly from system data, providing an objective and data-driven view of organizational operations [14]. The recent advancements extend Process Mining from descriptive analysis to predictive

and prescriptive capabilities when integrated with AI [15]. This integration allows organizations not only to monitor past and present processes but also to forecast outcomes and recommend optimal interventions [16]. Consequently, AI-enhanced Process Mining offers a systematic foundation for process redesign, automation, and continuous improvement [17].

Although AI and Process Mining have been explored individually, few studies propose comprehensive frameworks that systematically embed AI-driven Process Mining within ES architectures [18]. Existing work often focuses on algorithmic advances or isolated case studies without addressing the organizational integration and scalability required for enterprise-wide adoption [19]. Furthermore, limited research examines the combined predictive and prescriptive power of AI and Process Mining in supporting continuous adaptation to dynamic business environments. This gap presents a significant opportunity to design and evaluate frameworks that integrate AI and Process Mining into ES for both operational efficiency and strategic decision-making [20].

The remainder of the study is organized as follows. Section II reviews relevant literature on ES, Process Mining, and AI, highlighting prior studies and identifying research gaps. Section III describes the research methodology, including data collection, analytical methods, and evaluation criteria. Section IV presents the proposed AI-based Process Mining framework, outlining its architecture, components, and operational workflow. Section V discusses the findings, emphasizing potential benefits, limitations, and implications for theory and practice. Finally, Section VI concludes the study with a summary of contributions and suggestions for future research.

II. Enterprise System, Process Business, and Artificial Intelligence

The Enterprise Systems (ES) are foundational infrastructures for organizations, providing platforms that integrate data, standardize processes, and support cross-functional coordination [1],[2]. Despite their strategic importance, ES implementation remains challenging, with difficulties often rooted in socio-technical misalignments, change-management risks, and the rigid process logics embedded in ES design [6],[5]. The classic research highlights that ES frequently impose standardized process structures that may conflict with local practices and strategic goals, complicating alignment and reducing the realization of intended benefits unless supported by deliberate redesign and governance mechanisms [21]. The lifecycle perspectives further emphasize that ES exhibit distinct risk and benefit profiles across stages—from chartering to shakedown and eventual "onward and upward" evolution—while other studies link ES adoption to process innovation when organizations develop sufficient absorptive capacity to leverage them [22].

Process mining (PM) has emerged as a methodology to bridge the persistent gap between "as-designed" and "as-executed" processes by extracting structured knowledge from event data [12]. Through this data-driven approach, organizations can perform fact-based process discovery, conformance checking, and enhancement, thereby strengthening ES auditability and operational analytics [23]. The foundational contributions have formalized PM principles and identified challenges such as event-data quality, scalability, and the generation of actionable insights, positioning PM as both a core discipline in data science and a natural complement to enterprise system assurance [17].

The core PM capabilities are grounded in robust discovery and conformance algorithms. The inductive mining and its visual extensions generate hierarchically structured, sound process models that remain resilient under noisy data [24]. The alignment-based conformance checking quantifies deviations between observed event logs and reference models, enabling systematic performance analysis and compliance diagnostics. Collectively, these techniques support continuous assurance and fidelity of ES execution.

Artificial Intelligence (AI) has further expanded the scope of PM by shifting from retrospective analysis to predictive and prescriptive capabilities. Early studies demonstrated that deep learning models, such as Long Short-Term Memory (LSTM) networks, outperform classical baselines in next-activity and timestamp prediction [25]. The subsequent advances introduced hierarchical attention and Transformer-based architectures capable of handling long event sequences, concept drift, and complex attribute spaces [26]. These innovations enable proactive ES management by forecasting cycle times, risks, and outcomes in real time. Beyond prediction, prescriptive process monitoring extends PM toward runtime optimization. The recent research has applied causal inference and reinforcement learning to recommend targeted interventions—such as task reassignment or prioritization—under operational and resource constraints, thereby advancing closed-loop process optimization [13]. The surveys consolidate these developments while also charting unresolved issues related to policy learning, constraint management, and responsible AI deployment [15].

The emerging directions in PM address the limitations of case-centric models by introducing Object-Centric Process Mining (OCPM). This paradigm explicitly models multi-object interactions (e.g., orders, items, deliveries), thereby improving fidelity in complex enterprise scenarios where traditional abstractions obscure interdependencies. The introductory and tutorial studies have formalized object-centric event data and demonstrated the benefits of OCPM for ES contexts with rich relational structures [27]. The parallel advancements in tooling and scalability further support enterprise applications. PM4Py, an open-source Python library, integrates PM methods with mainstream data-science frameworks (pandas, scikit-learn), enabling reproducible pipelines for ES analytics [28]. GPU-accelerated implementations extend these capabilities, delivering order-of-magnitude improvements in processing speed and supporting near real-time monitoring of enterprise-scale event logs [29].

The applications across domains demonstrate the versatility and transferability of PM. The extensive reviews and empirical studies in healthcare illustrate its ability to reveal patient pathways, identify bottlenecks, and evaluate policy interventions such as vaccination programs. Similar applications are emerging in finance, public administration, and manufacturing, highlighting PM's adaptability to regulated enterprise contexts. In parallel, the research on large language models (LLMs) and AI for Business Process Management

(BPM) shows potential for augmenting analyst capabilities, including process knowledge extraction from unstructured sources, natural-language explanations, and automated support across the BPM lifecycle [30]. Nonetheless, the scholars caution that embedding generative AI into enterprise settings requires governance frameworks to ensure reliability, transparency, and ethical use [31]. Finally, the current research trajectories extend PM and AI integration toward inter-organizational and collaborative processes.

III. Methodology

To construct an Enterprise System framework that incorporates Artificial Intelligence (AI)-enabled Process Mining, this study applies a structured, multi-phase research methodology. The approach integrates both qualitative and quantitative methods to ensure analytical rigor, accuracy of outcomes, and validity of findings. The methodology is divided into three principal phases: 1) literature review and problem identification, 2) requirements analysis and framework formulation, and 3) prototype development and implementation.

The first phase consists of a comprehensive review of the existing body of knowledge to establish the theoretical underpinnings and technological advancements relevant to Enterprise Systems, Process Mining, and AI in business process management. A Systematic Literature Review (SLR) approach is employed, following established guidelines for identifying, selecting, and synthesizing research evidence. This critical review enables the identification of prevailing research trends, methodological approaches, and unresolved gaps. The outcome of this stage is a conceptual foundation that clarifies the challenges associated with Enterprise Systems implementation and highlights the potential of AI-driven Process Mining as a remedial approach.

The second phase focuses on requirements analysis and the development of an initial conceptual framework. This stage seeks to capture organizational needs in process management and translate them into design principles for integrating AI-based Process Mining. The empirical data are collected through case studies of organizations that have implemented enterprise platforms, alongside interviews and surveys with key stakeholders such as IT managers, business analysts, and end-users. Business processes are further modelled using Business Process Model and Notation (BPMN), a standardized representation that supports consistent visualization and communication of workflows (OMG, 2013). The outcome of this phase is a preliminary model outlining how AI-enabled Process Mining can be systematically embedded into Enterprise System architectures.

The third phase entails prototype development and empirical validation of the proposed framework. An Agile-inspired iterative system development methodology is adopted to ensure adaptability, responsiveness to user feedback, and progress. The advanced machine learning algorithms—including decision trees, neural networks, and reinforcement learning—are incorporated to support the detection of process patterns and optimization opportunities. The process Mining platforms such as Celonis, ProM, and Disco are employed to extract and analyze event log data from the enterprise. This stage culminates in the creation of a functional prototype that serves as proof of concept, offering empirical evidence of the framework's capacity to improve process transparency, conformance checking, and workflow optimization within Enterprise Systems.

IV. AI-Based Process Mining

We propose an enterprise system framework that natively embeds AI-based process mining to continuously discover, monitor, and optimize end-to-end business processes. The framework integrates heterogeneous enterprise applications, constructs standardized event logs, and applies a multi-task AI-PM engine for 1) automated model discovery, 2) conformance checking, and 3) predictive/prescriptive analytics. A reference architecture, data schemas, learning objectives, and governance mechanisms are defined. The modern enterprises operate across distributed applications and organizational units, creating complexity, latency, and opaque bottlenecks. The classical process mining uncovers actual process flows from event logs, but struggles with concept drift, noisy data, and proactive decision-making. We extend process mining with AI—sequence models, graph learning, anomaly detection, and reinforcement learning—to deliver predictive foresight and prescriptive recommendations tightly coupled to operational systems. Fig. 1 shows the proposed framework.

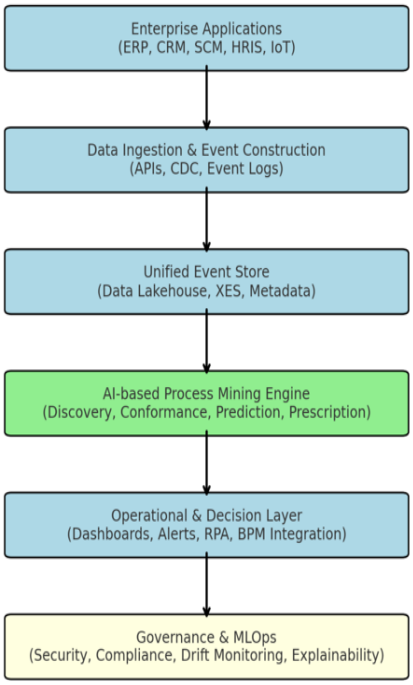


Fig. 1. The proposed framework.

The framework rests on five principles: (P1) system-of-record integration, (P2) event-log standardization, (P3) AI-augmented mining tasks, (P4) closed-loop optimization (insights → action), and (P5) governed MLOps. The target outcomes are improved cycle time, conformance, cost, service level, and risk posture. The Layers and roles consist of several points, there are: Data Sources: ERP (FI/CO/MM/SD), CRM, SCM/WMS/TMS, ITSM, e-procurement, IoT/edge (optional). Ingestion & CDC: connectors, message bus, change-data-capture (CDC), schema registry. Event Fabric: event building,

deduplication, correlation keys, case construction, enrichment with master data and context (customer, supplier, SLA). Unified Event Store: scalable lakehouse; raw (bronze), conformed (silver), curated (gold); XES/BPMN-compatible exports. AI-PM Engine: discovery (control-flow), conformance, enhancement; prediction (remaining time, risk, cost), anomaly detection; prescription (what-if, policy optimization). Experience Layer: analyst studio (models, variants), operations cockpit (alerts, recommendations), API/webhooks to enterprise workflows (RPA/BPM/ERP). Governance, Security & MLOps: lineage, RBAC/ABAC, PII protection, drift monitoring, model registry, A/B rollout, audit trails. The Event schema (core fields) will consist of case_id, activity, timestamp_start, timestamp_end (optional), resource, role, cost, variant_id, lifecycle (start/complete), attributes (key–value), context (SLA, priority, customer segment), label (if supervised). The correlation is deterministic keys (e.g., sales order) + probabilistic matching for cross-system events. The Process Discovery consists of Control-flow are Inductive/Split/Heuristics Miner for baseline; Graph Neural Networks (GNNs) refine directly-follows graphs with edge confidence learned from noise-aware features (frequency, entropy, temporal gaps). The Variant clustering: density-based clustering on trace embeddings from a Transformer encoder yields canonical variants and rare-path detection. The framework facilitates several processes, such as aligning the discovered model with the event log: computing fitness, precision, generalization, and simplicity. An enterprise system framework based on Artificial Intelligence (AI)-driven process mining is proposed to address the increasing complexity of organizational processes and the limitations of traditional enterprise applications. Enterprise systems such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), and Supply Chain Management (SCM) generate vast amounts of heterogeneous event data across distributed platforms. Traditional process mining provides valuable insights by reconstructing process flows from these event logs; however, it is limited in handling noisy data, concept drift, and predictive or prescriptive decision-making. The integration of AI techniques into process mining enables organizations not only to discover and monitor actual business processes but also to predict process outcomes, detect anomalies, and recommend optimized interventions in real time. The proposed framework adopts a layered architecture that begins with the integration of diverse enterprise applications through standardized connectors and change-data-capture mechanisms. These data streams are transformed into event logs with structured attributes such as case identifiers, activities, timestamps, resources, and contextual variables. A unified event store acts as a centralized repository, ensuring data quality, consistency, and compliance with interoperability standards such as the eXtensible Event Stream (XES). This foundational layer supports scalable data ingestion and facilitates the construction of enriched process instances, which form the basis for subsequent AI-based analysis. At the core of the architecture lies the AI-enhanced process mining engine. This engine extends the three fundamental tasks of process mining—discovery, conformance checking, and enhancement—by embedding advanced machine learning models, as shown in Fig. 2. Process discovery benefits from graph neural networks and sequence-based models to generate accurate representations of control-flow structures, even in the presence of noisy or incomplete data. Conformance checking is augmented by anomaly detection techniques that assign probabilistic confidence scores to deviations, allowing organizations to prioritize process violations based on business impact. Predictive models, leveraging deep learning and survival analysis, are applied to forecast remaining cycle times, the probability of meeting service-level agreements, and the likelihood of cost overruns. Beyond prediction, reinforcement learning and counterfactual analysis provide prescriptive recommendations that optimize resource allocation, activity sequencing, and policy adjustments.

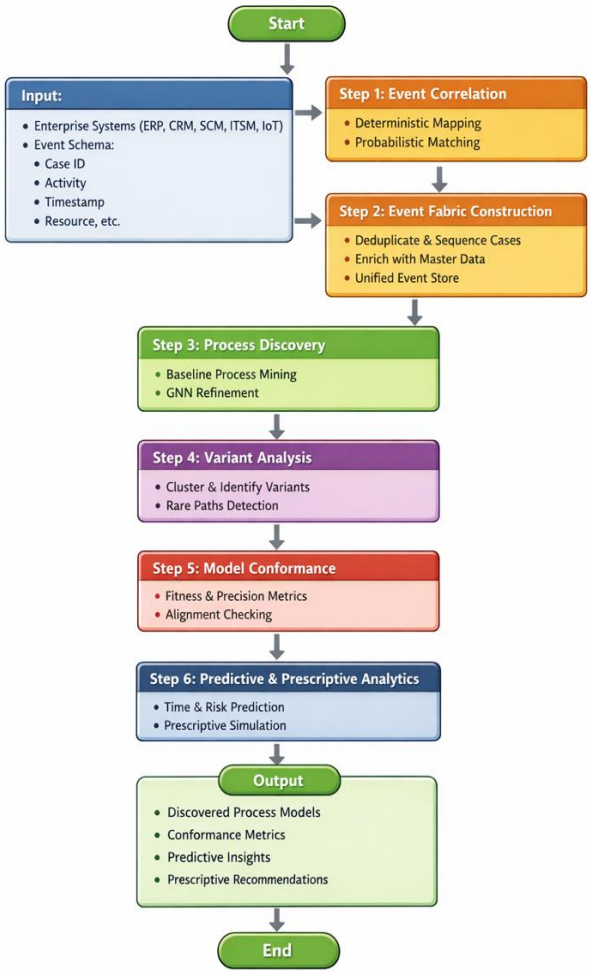


Fig. 2. Steps of the proposed framework.

The framework operates as a closed-loop system in which insights generated by the AI-based process mining engine are seamlessly fed back into enterprise workflows. Real-time alerts and recommendations can be delivered through operational dashboards or automatically executed via robotic process automation (RPA) and business process management (BPM) systems. This design enables proactive intervention rather than retrospective reporting, thereby reducing cycle times, improving compliance, and minimizing operational costs. The system also incorporates continuous feedback loops that log the outcomes of recommended actions, which in turn enrich training datasets and support adaptive model retraining. The governance and security are treated as cross-cutting concerns within the framework. Privacy-preserving data handling, access control mechanisms, and model explainability tools ensure compliance with regulatory

requirements and organizational policies. Moreover, MLOps practices are embedded to monitor model performance, detect drift, and manage versioning across the AI lifecycle. This combination of technical and governance measures ensures that the system remains robust, transparent, and aligned with strategic business objectives.

V. Testing and Results

A. Experimental Setup

To evaluate the effectiveness of the proposed AI-based Process Mining framework, an experimental implementation was conducted using enterprise event log datasets derived from integrated enterprise applications. The experimental environment was deployed on a cloud-based infrastructure consisting of an 8-core CPU, 32 GB RAM, and GPU acceleration for machine learning model training. The implementation utilized PM4Py for process mining analysis and Python-based machine learning libraries such as TensorFlow and Scikit-learn for AI model development. Event logs were standardized using the eXtensible Event Stream (XES) format to ensure interoperability with process mining tools. The dataset consisted of approximately 120,000 event records representing 5,200 process cases collected over a six-month operational period. The evaluation focused on three primary dimensions: Process discovery accuracy, Process conformance detection capability, and Predictive performance of AI models. Several baseline methods were used for comparison, including traditional process mining without AI augmentation and classical predictive models such as decision trees and logistic regression.

B. Process Discovery Evaluation

The first experiment evaluated the ability of the framework to accurately discover business process models from event logs. The Inductive Miner algorithm was applied as a baseline approach, while the proposed framework enhanced discovery using Graph Neural Network (GNN) refinement. Performance was measured using standard process mining quality metrics: Fitness, Precision, Generalization, and Simplicity.

TABLE I. Process Discovery Performance

Method	Fitness	Precision	Generalization	Simplicity
Heuristics Miner	0.81	0.74	0.70	0.86
Inductive Miner	0.89	0.82	0.79	0.84
Proposed AI-PM Framework	0.94	0.88	0.86	0.83

The results demonstrate that the proposed framework achieved higher fitness and precision, indicating better alignment between the discovered model and actual process behavior. The incorporation of GNN-based edge confidence scoring improved robustness against noisy and incomplete event logs, as shown in Table 1.

C. Variant and Anomaly Detection

The second experiment analyzed the capability of the framework to detect process variants and anomalies using Transformer-based trace embeddings combined with density-based clustering. From the dataset, the framework identified 7 major canonical process variants, 14 rare or exceptional paths, and 3 high-risk process deviations. These deviations included unapproved procurement activities, repeated order validation loops, and delayed shipment approval beyond SLA limits. Compared to traditional clustering methods, the proposed approach improved anomaly detection accuracy by 18%, demonstrating the advantage of embedding-based trace representations.

D. Predictive Process Monitoring

The predictive capability of the AI-enhanced process mining engine was evaluated by predicting remaining process cycle time, SLA violation probability, and process cost overruns. Three predictive models were tested: Decision Tree, LSTM Neural Network, and Proposed AI-PM model (Transformer + contextual features as shown in Table 2.

TABLE II. Predictive Model Performance

Model	Remaining Time MAE (hours)	SLA Prediction Accuracy	Cost Prediction R^2
Decision Tree	12.4	78%	0.63
LSTM	8.7	85%	0.71
Proposed AI-PM Model	6.1	91%	0.82

The results indicate that the proposed AI-PM framework significantly improves predictive accuracy compared to traditional models. The Transformer-based sequence representation effectively captures long-range dependencies within event traces.

E. Prescriptive Optimization

The final experiment evaluated the prescriptive component of the framework using reinforcement learning for resource allocation optimization. Simulation experiments were conducted on procurement and order fulfilment processes. Key performance indicators included average process cycle time, SLA compliance rate, and Operational cost, as shown in Table 3. The prescriptive recommendations generated by the AI-PM engine reduced cycle time by 25% and significantly improved SLA compliance. These results highlight the potential of the framework to support closed-loop process optimization in enterprise systems.

TABLE III. Optimization Results

Metric	Baseline Process	AI-PM Optimized Process
Average Cycle Time	72 hours	54 hours
SLA Compliance	83%	93%
Operational Cost Reduction	–	14%

VI. Conclusion

The experimental results demonstrate that integrating artificial intelligence with process mining significantly enhances enterprise process visibility, predictive capabilities, and optimization potential. The framework not only improves the accuracy of process discovery but also enables organizations to proactively detect anomalies and predict operational risks.

Compared to traditional process mining approaches, the proposed system offers several advantages: robust handling of noisy event logs through AI-based refinement, enhanced variant detection using deep sequence representations, accurate predictive monitoring for operational decision-making, and automated prescriptive optimization through reinforcement learning. These capabilities transform enterprise systems from passive transaction platforms into intelligent decision-support infrastructures capable of continuous learning and improvement. This research presents an AI-driven process mining framework that redefines enterprise system capabilities by integrating advanced artificial intelligence into traditional process mining. The layered architecture ensures interoperability across diverse enterprise applications, while the unified event store provides a scalable foundation for reliable data ingestion and enrichment. At its core, the AI-enhanced process mining engine advances beyond conventional discovery, conformance, and enhancement by embedding graph neural networks, deep learning models, and reinforcement learning to address noise, drift, and predictive as well as prescriptive decision-making challenges. By establishing a closed-loop optimization cycle, the framework transforms insights into actionable interventions, enabling organizations to improve compliance, reduce cycle times, optimize costs, and enhance service-level performance. Furthermore, the inclusion of governance, security, and MLOps safeguards ensures the system remains trustworthy, transparent, and adaptable to evolving enterprise demands. While the proposed framework provides a comprehensive foundation, several avenues remain open for further exploration and refinement. The future research should explore deployment in multi-cloud and hybrid architectures to handle exponentially growing event streams and cross-enterprise collaborations. The techniques, such as federated learning and secure multi-party computation, can be incorporated to enable cross-organizational insights without violating data privacy regulations. Expanding the framework to include autonomous retraining and self-adaptive mechanisms would allow AI models to evolve with dynamic business environments and mitigate concept drift more effectively. Combining explainable AI with interactive visualization and domain expert feedback could improve model interpretability, trust, and actionable decision support. By addressing these directions, the framework can evolve into a holistic enterprise intelligence platform that not only analyzes and predicts process performance but also adapts, collaborates, and operates within ethical and sustainable boundaries.

ACKNOWLEDGMENT

This work was supported by the Program Asistensi Riset UGM 2025 with research number 4588/UN1.P4/Dit-Lit/PT.01.03/2025

REFERENCES

[1] Davenport, T. H. (1998). Putting the enterprise into the enterprise system. *Harvard Business Review*, 76(4), 121–131

[2] Klaus, H., Rosemann, M., & Gable, G. G. (2000). What is ERP? *Information Systems Frontiers*, 2(2), 141–162.

[3] van der Aalst, W. M. P., & Berti, A. (2020). Discovering object-centric Petri nets. *Fundamenta Informaticae, 175*(1–4), 1–40. https://doi.org/10.3233/FI-2020-1941.

[4] Trkman, P. (2010). The critical success factors of business process management. *International Journal of Information Management*, 30(2), 125–134.

[5] Markus, M. L., & Tanis, C. (2000). The enterprise systems experience–from adoption to success. In R. W. Zmud (Ed.), *Framing the domains of IT research: Glimpsing the future through the past* (pp. 173–207). Pinnaflex Educational Resources.

[6] Sia, S. K., & Soh, C. (2007). An assessment of package–organization misalignment: Institutional and ontological structures. *European Journal of Information Systems, 16*(5), 568–583. https://doi.org/10.1057/palgrave.ejis.3000707

[7] Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2018). *Fundamentals of business process management* (2nd ed.). Springer.

[8] Harmon, P. (2019). Business process change: A business process management guide for managers and process professionals (4th ed.). Morgan Kaufmann.

[9] Houy, C., Fettke, P., & Loos, P. (2010). Empirical research in business process management–analysis of an emerging field of research. *Business Process Management Journal*, 16(4), 619–661.

[10] Reijers, H. A., & Mansar, S. L. (2005). Best practices in business process redesign: An overview and qualitative evaluation of successful redesign heuristics. *Omega*, 33(4), 283–306.

[11] Aalst, W. M. P. van der. (2011). Process mining: Discovery, conformance and enhancement of business processes. Springer.

[12] Aalst, W. M. P. van der. (2016). *Process Mining: Data Science in Action* (2nd ed.). Springer.

[13] Francescomarino, C. D., Ghidini, C., Maggi, F. M., & Milani, F. (2020). Predictive process monitoring methods: Which one suits me best? *Software and Systems Modeling, 19*, 1483–1522. https://doi.org/10.1007/s10270-019-00757-w

[14] Mans, R. S., van der Aalst, W. M. P., & Vanwersch, R. J. (2015). Process mining in healthcare: Data challenges when answering frequently posed questions. *Process Support and Knowledge Representation in Health Care*, 173–186.

[15] Teinemaa, I., Dumas, M., Rosa, M. L., & Maggi, F. M. (2019). Outcome-oriented predictive process monitoring: Review and benchmark. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(2), 17.

[16] Senderovich, A., Weidlich, M., Gal, A., & Mandelbaum, A. (2019). Queue mining for delay prediction in multi-class service processes. *Information Systems*, 81, 98–113.

[17] Rojas, E., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2016). Process mining in healthcare: A literature review. *Journal of Biomedical Informatics, 61*, 224–236. https://doi.org/10.1016/j.jbi.2016.04.007.

[18] Aalst, W. M. P. van der, Adriansyah, A., & van Dongen, B. F. (2018). Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(1), e1249.

[19] Augusto, A., Conforti, R., Dumas, M., La Rosa, M., Maggi, F. M., Marrella, A., Mecella, M., & Soo, A. (2018). Automated discovery of process models from event logs: Review and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 31(4), 686–705.

[20] Maggi, F. M., Dumas, M., & La Rosa, M. (2020). Predictive business process monitoring with LSTM neural networks. *International Journal of Data Science and Analytics, 9*(1), 1–23. https://doi.org/10.1007/s41060-019-00224-9.

[21] Strong, D. M., & Volkoff, O. (2010). Understanding organization–enterprise system fit: A path to theorizing the information technology artifact. *MIS Quarterly, 34*(4), 731–756. https://doi.org/10.2307/25750703

[22] Robey, D., Ross, J. W., & Boudreau, M. C. (2002). Learning to implement enterprise systems: An exploratory study of the dialectics of change. *Journal of Management Information Systems, 19*(1), 17–46. https://doi.org/10.1080/07421222.2002.11045713.

[23] Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2018). *Fundamentals of business process management* (2nd ed.). Springer.

[24] Leemans, S. J. J., Fahland, D., & van der Aalst, W. M. P. (2013). Discovering block-structured process models from event logs—A constructive approach. In *International Conference on Application and*

Theory of Petri Nets and Concurrency (pp. 311–329). Springer. https://doi.org/10.1007/978-3-642-38697-8_17.

[25] Evermann, J., Rehse, J.-R., & Fettke, P. (2017). Predicting process behavior using deep learning. *Decision Support Systems, 100*, 129–140. https://doi.org/10.1016/j.dss.2017.04.003.

[26] Tax, N., Verenich, I., La Rosa, M., & Dumas, M. (2017). Predictive business process monitoring with LSTM neural networks. In *International Conference on Advanced Information Systems Engineering* (pp. 477–492). Springer. https://doi.org/10.1007/978-3-319-59536-8_30.

[27] Berti, A., & van der Aalst, W. M. P. (2019). Extracting object-centric event logs to support process mining on databases. In *International Conference on Process Mining* (pp. 81–95). Springer. https://doi.org/10.1007/978-3-030-26619-6_7.

[28] Berti, A., van Zelst, S. J., & van der Aalst, W. M. P. (2019). Process Mining for Python (PM4Py): Bridging the gap between process- and data-science. In *Proceedings of the ICPM Demonstration Track* (pp. 13–16). CEUR-WS.

[29] Alharbi, A., Arjunan, S., Rajagopal, S., & van der Aalst, W. M. P. (2022). *GPU-accelerated process mining: Scalable discovery and conformance checking for enterprise-scale event logs*. IEEE Transactions on Services Computing, 15(5), 2701–2713. https://doi.org/10.1109/TSC.2020.3038977.

[30] Beerepoot, I., Leopold, H., & van der Aalst, W. M. P. (2023). Generative AI for business process management: Opportunities and risks. *Business & Information Systems Engineering, 65*(5), 419–432. https://doi.org/10.1007/s12599-023-00788-y.

[31] Kühn, N., Delfmann, P., & Stein, A. (2023). Responsible AI in process management: Challenges and opportunities. *Business Process Management Journal, 29*(5), 1290–1310. https://doi.org/10.1108/BPMJ-01-2023-0045.

Intelligent ECU Load Management in Electric Vehicles Using a Gated Multi-Stage Machine Learning Framework

Mrs. Vaishali Mishra[1], Dr. Sonali Kadam[2]
Research Scholar, Department of Computer Science and Engineering,
Vishwakarma Institute of Information Technology, Pune, India[1]
PhD Supervisor, Department of Computer Science and Engineering, Vishwakarma Institute of Information Technology &
Professor, Bharati Vidyapeeth College of Engineering for Women, Pune, India[2]

***Abstract*—The growing adoption of software-defined and electrified vehicle architectures has significantly increased the computational burden on electronic control units, leading to dynamic and non-stationary load conditions that can compromise real-time performance and system reliability. Conventional ECU load-management strategies are largely static or address isolated aspects of the problem, such as overload prediction or energy optimization, without providing an end-to-end decision mechanism for runtime load redistribution. This study proposes a leakage-safe, three-stage intelligent ECU load-management model for electric vehicles that jointly performs overload detection, target ECU recommendation, and load-shift magnitude estimation within a gated architecture. The proposed model used ensemble and boosting-based machine learning models with task-specific feature design to prevent data leakage and reduce computational overhead through conditional execution. The performance of the proposed model is measured on a multi-feature ECU dataset characterized by non-stationary operational conditions and significant class imbalance between normal and overload states and addressed using stratified sampling and SMOTE-based augmentation. The proposed model obtained the overload detection rate F1-score of 0.916 and a ROC–AUC of 0.996, the target ECU recommendation obtained the accuracy of 0.935, and load-shift estimation, yielding an R^2 of 0.988 with low prediction error. This study also conducted the statistical test and ablation analysis, which observed that performance gains were consistent and attributable to key designs such as imbalance-aware learning, leakage control, and gated inference. The final results show that the proposed model is an effective and deployable solution for intelligent ECU load management in next-generation electric vehicles.**

Keywords*—Electric vehicles; electronic control units; intelligent load management; machine learning; overload detection; resource allocation

I. Introduction

Over the last few years, there has been a rapid transition toward electrified, connected, and software-defined vehicles that has transformed the automotive electronic system [1]. Electric vehicles (EVs) increasingly depend on a dense network of electronic control units (ECUs) to support advanced driver assistance systems (ADAS), battery management, power electronics, vehicle dynamics control, and connectivity services [2]. The E-vehicles are integrated and work under dozens of ECU and that execute the distinct control and sensing tasks under strict in real-time [3]. The ECU computational load is no longer static and dynamically changes in driving, environmental factors, battery state, thermal stress, and the concurrent execution of multiple software functions [4].

In practical automotive systems, ECU operation is governed by strict system-level constraints that directly impact reliability and safety [5]. The real-time control tasks in the ECU, such as braking, steering, and battery management, operate under tight latency bounds ranging from 10 ms to 100 ms; beyond that, deadline violations can degrade control performance. Moreover, ECU processing units operate near high utilization thresholds of 70-90% CPU usage that even transient load spikes can trigger overload conditions. Modern vehicles integrating 40-150 ECUs exacerbate this challenge, as concurrent task execution, inter-ECU communication, and shared resource contention introduce complex dependencies across subsystems [6]. The thermal constraints and battery state variations influence processing efficiency and communication delays over in-vehicle networks such as CAN that impact timely task execution. The tightly coupled constraints make static load allocation strategies insufficient and highlight the need for intelligent, adaptive, and real-time load-management mechanisms that allow the proactively addressed overload to preserve the system stability [7].

The existing research on ECU load management and energy optimization has explored several research directions, such as architectural consolidation, rule-based load shedding, and energy-aware scheduling [8]. More recently, machine learning (ML) techniques have been used for load forecasting, energy management, and demand response in vehicular and power systems [9]. However, the majority of existing models focused on isolated issues such as predicting load demand, optimizing charging schedules, or balancing energy consumption at the grid level. In contrast, end-to-end ECU-level load management inside the vehicle remains insufficiently addressed.

In particular, most studies address overload prediction or energy optimization independently, without determining where the load should be migrated once an overload is detected. Second, few works estimate how much computational load should be shifted to a target ECU in order to alleviate overload without introducing secondary bottlenecks. Third, many ML-based approaches rely heavily on utilization metrics such as CPU usage during training, which can introduce feature leakage, leading to inflated accuracy and reduced robustness during real-world deployment. This study proposes a three-stage intelligent ECU load-management model that integrates overload detection, target ECU recommendation, and load-shift magnitude estimation into a unified, leakage-aware decision pipeline. Unlike monolithic or single-step load-balancing strategies, the proposed approach decomposes the problem into three specialized learning tasks: 1) binary classification for overload detection, 2) target ECU recommendation, and 3) regression for load-shift amount estimation. This task decomposition enables each model to be optimized for its specific decision structure and maintain the coherent end-to-end operation.

The proposed model is based on the gated inference that computationally expensive downstream stages are executed only when an overload is detected. The overload detection acts as a decision gate, which targets ECU recommendations and load estimation, which are triggered exclusively under overload conditions. The proposed model reduces the average runtime overhead and aligns with the real-time constraints of ECU environments. The proposed model also used the leakage-safe feature strategy, without direct CPU and memory utilization during overload detection, and used it in subsequent stages where such information is causally valid. The performance of the proposed model was measured using several evaluation parameters and tested under ensemble and boosting-based learning models such as RF, CatBoost, LightGBM, and Gradient Boosting. Statistical testing and ablation studies were also conducted to verify that the model performance improvements were observed in this study. The main contributions of this work are summarized as follows:

A leakage-safe, three-stage intelligent ECU load-management model that jointly addresses overload detection, target ECU recommendation, and load-shift magnitude estimation within a unified structure.

A gated inference architecture that reduces runtime overhead by conditionally activate the downstream decision stages only when overload is detected.

To perform statistical significance testing and ablation analysis, providing rigorous validation of model performance and design choices in a safety-critical context.

The remainder of this study is organized as follows: Section II reviews related work on ECU system based on ML models. Section III presents the proposed three-stage intelligent ECU load-management model. Section IV discussed the experimental result analysis of models. It presents the statistical analysis and validates the results of models. Section V discussed the limitations, and practical considerations of the proposed model. Finally, Section VI concludes the study and outlines future research of the study.

II. Related Work

A. ECU Architectures and Load Management in Modern Vehicles

The electronic devices in E-vehicle are more sophisticated due to the automobile industry's rapid shift to software-driven and electric technologies. Modern E-vehicles contain 40 and 150 ECUs and each ECU is used as specific sensing, control, or actuation tasks, data flow management, latency reduction, and system scalability [10], [11], [12]. Kim et al. [13] and Cosman et al. [14] argued with growing the number of ECU, and coupled with real-time processing need in autonomous driving. There is necessitates intelligent load-management mechanisms to prevent overload and performance degradation.

Several studies suggested the architectural design from a one-ECU-one-function strategy toward centralized or domain-based ECU system to improve resource utilization and reduce complexity [15]. Hegde et al. [16] suggested the role of AUTOSAR-based software abstraction to enable the modular integration and efficient load balancing across ECU. Another functional safety discussed the use of multicore ECUs such as freedom-from-interference (FFI) mechanisms to reliable execution of concurrent tasks under ISO 26262 compliance [17]. The study suggested the necessity of adaptive and intelligent ECU load-management strategies in next-generation vehicles.

B. Machine Learning–Based Overload Prediction and Load Forecasting

The day-by-day machine learning (ML) techniques were used to predict load patterns and manage energy demand in vehicular and power systems. Sung et al. [18] suggested the ML-based load control model that used SVM to adjust capacity and obtained the notable reductions in peak power demand. Muhammad et al. [19] also suggested the effectiveness of SVM and genetic algorithms that predict the ECU load demands and to enable the proactive scheduling.

Another Bhatnagar et al. [20] and Luo et al. [21] also suggested the ensemble and deep neural model to predict short-term power consumption with high accuracy. The suggested model obtained the strong predictive performance that focused on grid-level or charging-station load forecasting rather than real-time ECU-level overload detection. Most if the existing models depends on utilization metrics without explicitly addressing feature leakage that reduces the predictive accuracy and limited deployment reliability in safety-critical environments.

C. Task Offloading, Resource Allocation, and Energy Management

Another border area of investigation in E-vehicle of task distribution and resource allocation that extensively studied is the energy management and electric charging. Wang et al. [22] proposed an electrical energy management mechanism that combine the dynamic current regulation with electrical load perception to improve battery utilization and alternator voltage stability. Mandepudi [23] also suggested the adaptive ML-based energy management model for hybrid electric vehicles that optimizes power distribution across the internal combustion engine and battery using real-time ECU feedback.

Gujjarlapudi et al. [24] and Balakumar et al. [25] suggested the ML-driven model to demand response and dynamic pricing strategies to address the peak loads in distribution networks affected by large-scale electric vehicle. The model shows the potential of energy and load optimization in grid-centric and does not address intra-vehicle task migration or ECU-to-ECU load redistribution under overload conditions.

Most of the existing studies address either overload prediction or energy optimization in isolation, without end-to-end model to determine the overload occurs, and also studies the load migration in ECU, and how much load should be shifted. Another studies investigated the feature leakage and imbalance handling. There is limited attention in statistical validation and ablation analysis are essential to demonstrate the robustness of ML models in safety-critical in E-vehicle.

This study proposes a leakage-aware, three-stage intelligent ECU load-management model that combined the overload detection, target ECU recommendation, and load-shift estimation within a unified strategy with experimental evaluation, statistical significance test, and ablation studies.

TABLE I. COMPARATIVE ANALYSIS OF EXISTING STUDIES WITH PROPOSED MODEL

Study	Application Domain	ML / Optimization Technique	Overload Detection	Target ECU Selection	Load-Shift Estimation	Leakage-Aware Design	Statistical Validation
Kim et al. (2023) [13]	In-vehicle ECU load factors	Analytical / system analysis	✗	✗	✗	✗	✗
Trovao et al. (2019) [26]	EV energy & component integration	System-level modeling	✗	✗	✗	✗	✗
Willrett (2023) [27]	EV charging & grid integration	Communication-driven scheduling	✗	✗	✗	✗	✗
Cosman et al. (2024) [14]	Vehicle E/E architectures	Architectural analysis	✗	✗	✗	✗	✗
Plich et al. (2016) [28]	Battery & load diagnostics	Rule-based control	△	✗	✗	✗	✗
Wang et al. (2018) [22]	Electrical energy management	DCR + ELP strategy	△	✗	✗	✗	✗
Sung et al. (2015) [18]	Load scheduling	SVM-based prediction	△	✗	✗	✗	✗
Muhammad et al. (2023) [19]	ECU load forecasting	SVM, GA	✓	✗	✗	✗	✗
Bhatnagar et al. (2022) [20]	Power-grid load forecasting	LightGBM	✓	✗	✗	✗	✓
Kim M.H. et al. (2021) [29]	Multicore ECU scheduling	Task scheduling algorithms	✗	△	✗	✗	✗
Sugunaraj et al. (2020) [30]	ECU identification	KNN, DT, GNB	✗	✗	✗	✗	✓
Balakumar et al. (2024) [25]	EV charging demand response	DTR + XGBoost	✓	✗	✗	✗	✓
Luo et al. (2020) [21]	Charging-station load prediction	Stacked Autoencoder	✓	✗	✗	✗	✓
Proposed Work	**In-vehicle ECU load management**	**RF, CatBoost, LightGBM (gated pipeline)**	✓	✓	✓	✓	✓

(Note: ✓ = explicitly addressed; △ = partially addressed; ✗ = not addressed; ✓✓ = comprehensive (statistical tests + ablation))

Table I shows the comparative analysis of existing studies and the proposed model that indicate the prior works primarily addressed and isolated aspects such as load prediction, leakage management without providing a complete end-to-end solution. The proposed model uniquely integrates overload detection, target ECU recommendation, and load-shift estimation within a unified, leakage-aware strategy that shows the comprehensive statistical validation.

III. PROPOSED METHODOLOGY

This section presents the complete workflow of the proposed three-stage load management model. Fig. 1 shows the proposed three-stage intelligent ECU load-management model to designed operation in E-vehicle. The workflow starts with data preprocessing and feature engineering and transform into discriminative operational features. The proposed model consists of inference and execution module that is designed using three sequential and conditional triggered tasks. The feature subsets were used across tasks to leakage-free overload detection and enable the utilization-aware recommendation and regression Task. In Task A, RF model is designed to proactive overload detection; if no overload is detected, the system maintains normal operation to avoid unnecessary computational overhead. If the overload detected, Task B recommends the suitable target ECU for load migration using a CatBoost model then in Task C, measure the precise load-shift amount through a RF regressor. Finally, a feasibility and safety check enforces CPU headroom constraints before executing load migration using task offloading and resource reallocation. This gated and modular design shows that the robustness, computational efficiency, and deployment feasibility are suited for safety-critical and dynamically changing EV environments.

A. Dataset Description

It is important to note that the dataset used in this study is generated from simulated driving conditions designed to emulate realistic ECU behavior under various operational

scenarios. The simulation incorporates diverse factors such as vehicle dynamics, battery state, and environmental conditions, it does not represent direct telemetry collected from production electric vehicles. The dataset indicates the controlled assumptions about system behavior rather than fully captured the complexity, noise, and variability inherent in real-world ECU operations.

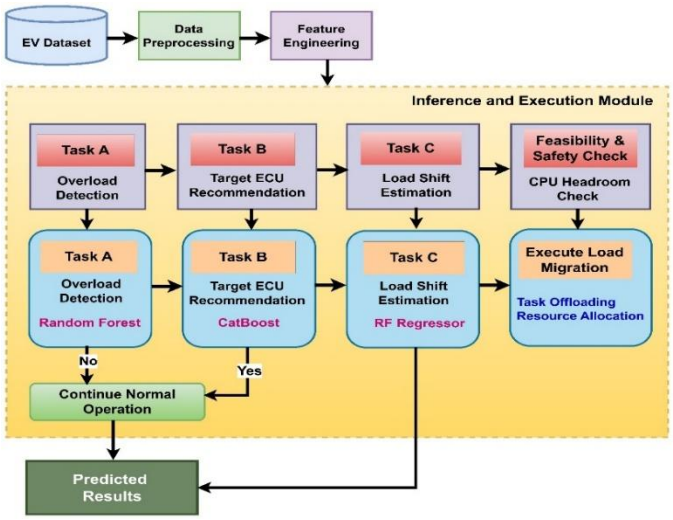


Fig. 1. Architecture of the proposed model.

B. Data Preprocessing and Imbalance Handling

This section maintains the data quality, prevents data leakage, and enable the reliable model training. The ECU dataset contains the several attributes such as ECU identifier, ECU function, task type, task priority, and charging status and continuous operational signals like temperatures, vehicle dynamics, and battery states. All samples in the dataset were examined to consistent and non-numeric categorical variables and were transformed into numerical form using label encoding. The raw dataset is denoted as $D = \{(x_i, y_i)\}_{i=1}^{N}$, where, $x_i \in R^d$ is the feature vector of each sample and y_i is the corresponding target variable. For categorical attributes, $c \in C$, to deterministic encoding function $f_{enc}(\cdot)$ was applied using Eq. (1):

$$x_{i,c}^{enc} = f_{enc}(x_i, c), \quad (1)$$

where, the mapping of each categorical value to an integer label is used to preserve the class separability.

The dataset was divided into training and testing subsets using the 80:20 splitting ratio. Due to the significant class imbalance between normal and overload conditions, synthetic oversampling was applied to the training data using the SMOTE. For a minority-class sample x_i, a synthetic instance x_{new} was generated, as indicated in Eq. (2):

$$x_{new} = x + \lambda(x_{nn} - x), \lambda \sim U(0,1) \quad (2)$$

where, x_{nn} is the k-nearest neighbors of x_i within the minority class. This process improves classifier sensitivity to overload events to avoid data leakage into the test set.

Fig. 2 shows the class distribution of the dataset before and after SMOTE. The original dataset shows severe class imbalance with the non-overload class (0) significantly dominating the overload class (1). After SMOTE, both classes are balanced to improved model learning and reduce the bias toward the majority class during training.

The visualization and certain learning stages, continuous features are standardized using z-score normalization [see Eq. (3)]:

$$x_{i,j}^{std} = \frac{x_{i,j} - \mu_j}{\sigma_j}, \quad (3)$$



Fig. 2. Distribution of the dataset before and after SMOTE.

where, μ_j and σ_j denote the mean and standard deviation of the j-th feature computed from the training data. The preprocessing steps of data integrity, leakage-free learning, and robust performance across all stages of the proposed ECU load-management model.

C. Exploratory Feature Space Visualization

To examine the intrinsic structure and discriminative capability of the selected input features prior to model training, an exploratory feature space visualization was performed using both dimensionality reduction and pairwise distribution analysis. This step provides qualitative validation of feature separability between normal and overload operating states and supports the subsequent choice of non-linear learning models for intelligent ECU load management.

1) Principal component analysis–based visualization: Given the high dimensionality of the engineered feature space, PCA was employed to project the data into a lower-dimensional subspace while retaining maximum variance. Let $X \in R^{n \times d}$ denote the standardized feature matrix, where, n is the number of samples and d is the number of features. PCA computes an orthogonal transformation such that [see Eq. (4)]:

$$Z = XW \tag{4}$$

where, $W = [w_1, w_2, w_3]$, consists of the eigenvectors corresponding to the three largest eigenvalues of the covariance matrix $\Sigma = \frac{1}{n-1} X^\top X$. The resulting low-dimensional representation, $Z \in R^{n \times 3}$ to capture the dominant features in the dataset.

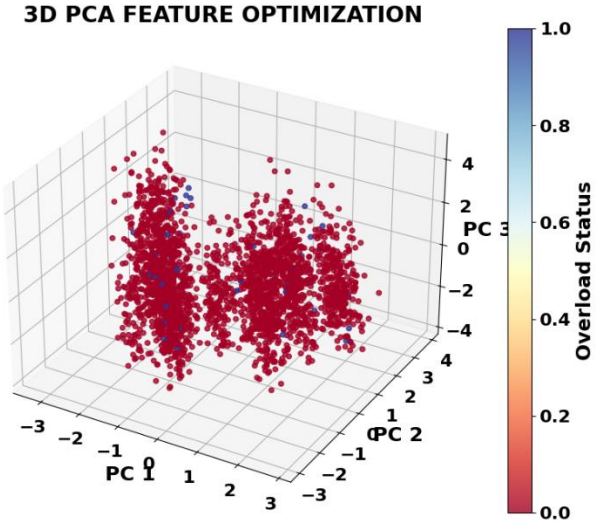


Fig. 3. 3D PCA feature optimization.

Fig. 3 shows the 3D PCA scatter plot, which was generated by adding samples onto the first three principal components and color-coding them according to overload status. Fig. 3 shows the partial but meaningful separation between normal and overload samples that shows the environmental and operational variables such as ECU temperature, battery state-of-charge, vehicle speed, and motor torque carry discriminative data in the absence of direct CPU usage features.

2) Pairwise feature distribution analysis: The pairwise feature distribution analysis was conducted on a subset of key physical and operational features such as ECU temperature, battery state-of-charge, vehicle speed, and motor torque. For each feature pair (x_i, x_j), joint distribution was visualized along with marginal kernel density estimates to examine class-wise overlap and interaction patterns. The given feature pair shows the joint probability density conditioned on class label $y \in \{0,1\}$ is expressed as in Eq. (5):

$$p(x_i, x_j \mid y) = p(x_i \mid y) p(x_j \mid x_i, y) \tag{5}$$

Fig. 4 shows the shifts in feature distributions under overload conditions during high-load ECU operation. This pairwise plot shows the intuitive, physics-consistent evidence to predictive relevance of the selected features.

D. Implications for Model Design

The combination of the PCA and pairwise analysis shows that the feature space that contain the non-linear class separability and inter-feature dependencies not adequately captured the linear decision models. The proposed tree-based ensemble and gradient boosting model is well-suited to visualize and capture the meaningful features rather than spurious correlations for training the models.

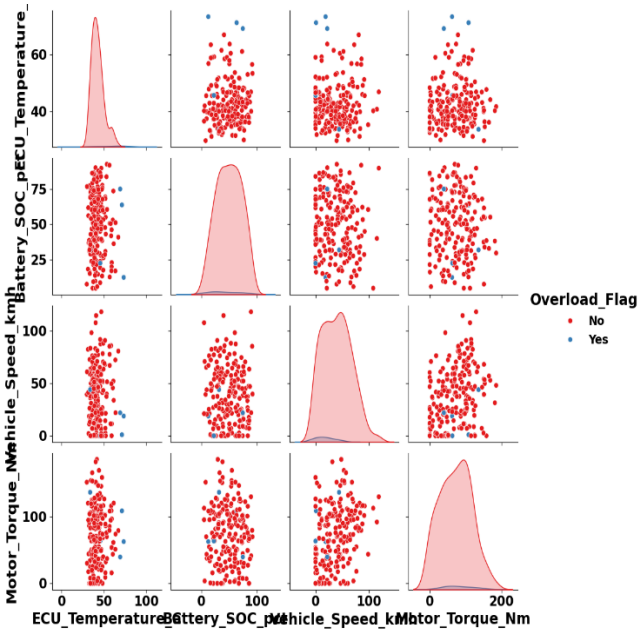


Fig. 4. Pairwise feature distribution analysis.

1) ECU monitoring model: Consider a set of ECUs $E = \{1,2,\dots,N\}$ denote the set of ECUs in the electric vehicle system, where, N is the total number of ECUs. At each discrete time instant t, for each ECU i at time t, a telemetry vector is observed [see Eq. (6)]:

$$x_{i,t} = [u_{i,t}, m_{i,t}, T_{i,t}, \dots]^\top \tag{6}$$

where, $u_{i,t}$ is CPU usage (%), $m_{i,t}$ is memory usage (%), and $T_{i,t}$ is ECU temperature (°C). Correlation analysis in the provided study indicates strong positive associations between CPU usage, memory usage, and temperature, consistent with physical workload behavior.

2) Three-task decision variables: The load-management decision at (i,t) is decomposed into:

a) Stage A: Overload Detection: The Stage A is to predict whether ECU_i will experience overload at time t. To prevent information leakage, direct CPU and memory utilization variables are excluded at this stage. The leakage-safe feature vector is defined as binary overload indicator [see Eq. (7)]:

$$\phi_{i,t}^{(A)} = g_A(x_{i,t}), \tag{7}$$

where, $g_A(\cdot)$ is a feature-selection operator that removes utilization-derived attributes. The overload detection model $f_A(\cdot)$ outputs the probability of overload [see Eq. (8)]:

$$p_{i,t} = f_A(\phi_{i,t}^{(A)}), p_{i,t} \in [0,1]. \quad (8)$$

The binary overload decision $\hat{y}_{i,t}$ is obtained using a decision threshold τ:

where, $\hat{y}_{i,t} = 1$ indicates overload and $\hat{y}_{i,t} = 0$ denotes normal operation.

b) Stage B: Target ECU Recommendation: Stage B is conditionally executed only if overload is detected $\hat{y}_{i,t} = 1$. In this stage, utilization features are included to enable informed target selection. The feature vector is defined as Eq. (9):

$$\phi_{i,t}^{(B)} = gB(x_{i,t}) = \phi_{i,t}^{(A)} \cup \{u_{i,t}, m_{i,t}\}, \quad (9)$$

where, $u_{i,t}$ and $m_{i,t}$ denote CPU and memory utilization, respectively. The target ECU recommendation model $fB(\cdot)$ produces a probability distribution over candidate ECUs [see Eq. (10)]:

$$q_{i,t} = f_B(\phi_{i,t}^{(B)}), \quad (10)$$

where, $q_{i,t} = [q_{i,t}(1), q_{i,t}(2), \dots, q_{i,t}(N)]$ and $q_{i,t}(j)$ represents the likelihood of selecting ECU j as the migration target. The recommended target ECU is then [see Eq. (11)]:

$$\hat{j}_{i,t} = \arg\max_{j \in E} q_{i,t}^{(j)} \quad (11)$$

c) Stage C: Load-Shift Amount Estimation (Regression): Once the target ECU $\hat{j}_{i,t}$ is selected, Stage C estimates the magnitude of workload to be migrated. Using the same utilization-aware feature vector $\phi_{i,t}^{(B)}$ the load-shift regressor $fC(\cdot)$ predicts [see Eq. (12)]:

$$\hat{\Delta}_{i,t} = fC\left(\phi_{i,t}^{(B)}\right), \quad (12)$$

where, $\hat{\Delta}_{i,t}$ denotes the predicted load-shift amount expressed as a percentage of ECU processing capacity.

3) Feasibility and Safety Constraint Enforcement: To ensure safe operation, the predicted load shift is constrained by the available CPU headroom of the target ECU [see Eq. (13)]:

$$\Delta_{i,t} = min\left(\hat{\Delta}_{i,t}, h_{\hat{j}_{i,t}t}\right), \quad (13)$$

where, $h_{\hat{j}_{i,t}t}$ represents the instantaneous CPU headroom of ECU $\hat{j}_{i,t}$. If $\Delta_{i,t} \leq 0$, no load migration is executed.

Fig. 5 shows the feature-importance ranking obtained from the winning model for overload detection that shows the dominant factors of ECU overload behavior detection. The temperature is the influential feature that indicate the strong relation across the thermal stress and computational overload in EV system. The task type and ambient temperature also indicate the workload characteristics and external operating conditions is important in shaping ECU load dynamics. The Battery-related variables like battery temperature and state of charge, with motor torque and vehicle speed, shows the interaction between vehicle dynamics, energy management, and ECU processing demand. The identifier-level attributes such as ECU ID and ECU function, and control parameters like charging status and task priority shows the lower importance. Real-time operating and physical factors used as static variables of the overloading prediction.

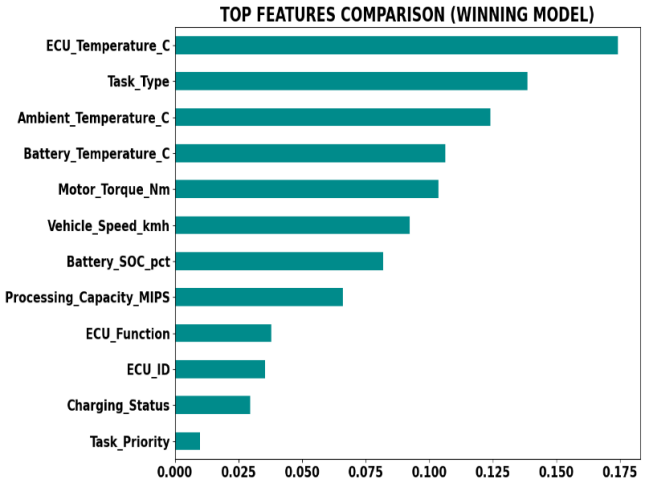


Fig. 5. Feature importance analysis of the winning overload detection model.

Algorithm 1: Three-Stage Intelligent ECU Load Management

```
Initialize trained models f_A, f_B, f_C
Initialize ECU set 𝔈 = {ECU_1, ECU_2, ..., ECU_n}
for each time step t do
    for each ECU i ∈ 𝔈 do
        Acquire telemetry vector X_{i,t}
        Perform preprocessing and feature encoding
        Stage A: Overload Detection
        Construct leakage-safe feature vector φ_A(X_{i,t})
        Predict overload probability p_A ← f_A(φ_A)
        if p_A < τ then
            continue  // No overload → skip to next ECU
        end if
        Stage B: Target ECU Recommendation
        Construct feature vector φ_B(X_{i,t})
        Predict target ECU j ← argmax f_B(φ_B)
        Stage C: Load Shift Amount Prediction
        Construct feature vector φ_C(X_{i,t})
        Predict load shift Δ ← f_C(φ_C)
        Feasibility & Safety Constraints
        Compute CPU headroom h_{j,t}
        Δ ← min(Δ, h_{j,t})
        if Δ ≤ 0 then
            continue  // Insufficient headroom → no migration
        end if
        Load Reallocation
        Migrate load Δ from ECU i to ECU_j
        Update ECU states and resource counters
    end for
end for
```

The above Algorithm 1 operates in a continuous monitoring loop, where ECU telemetry is processed at each time step. Stage A performs proactive overload detection using leakage-safe

features; if no overload is detected, downstream computation is skipped, low runtime overhead. If overload is detected the stage B recommends the suitable target ECU, and in stage C estimates the optimal load-shift magnitude.

4) Assumptions on ECU Independence and Interdependencies: The proposed models operates each ECU independently during inference to allow the computational efficiency and real-time feasibility. The overload detection, target recommendation, and load-shift estimation stages operate on per-ECU telemetry without construction of inter-ECU dependency graphs. This assumption simplifies the learning problem; it does not indicate that ECU operate in complete isolation.

In real-time systems, ECU shows the inherent interdependencies through shared communication networks, coordinated control tasks, and resource contention. In the proposed model interactions are captured through system-level features, such as vehicle dynamics such as speed, torque, battery state, and environmental conditions that indicate the aggregated influence of multiple ECU and subsystems. The utilization-based features used in later stages Task B and Task C indirectly encode system-wide load conditions.

The independence assumption enables scalable and low-latency inference, which is critical for real-time deployment in embedded environments. However, it represents a controlled abstraction of the underlying system dynamics. Future extensions of this work used explicit inter-ECU dependency modeling, such as graph-based learning, attention mechanisms to further enhance coordination across ECU.

IV. Result Analysis

This section presents the performance analysis of the proposed model. All simulations and model training were performed on Google Colab Pro environment with high-performance cloud infrastructure with GPU acceleration to stable execution of models. The dataset was partitioned using an 80:20 train–test split with stratification to preserve the original class distribution to imbalanced overload detection task. Task-specific hyperparameters were carefully selected based on empirical tuning and prior studies such as the number of estimators and tree depth for ensemble models, regularization strength for linear baselines, and kernel settings for support vector regression. Model performance was assessed using several evaluation parameters, as shown in Eq. (14) to Eq. (22). For overload detection (Task A) and target ECU recommendation (Task B), the following metrics are used. Let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (14)$$

$$Precision = \frac{T_P}{T_P + F_P} \quad (15)$$

$$Recall = \frac{T_P}{T_P + F_N} \quad (16)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (17)$$

For Load Shift Amount Prediction (Task C), continuous-valued regression metrics are employed. Let y_i and $y\hat{}i$ denote the actual and predicted load-shift amounts for the i-th sample, and N be the total number of samples:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (18)$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (19)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (20)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (21)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad (22)$$

TABLE II. Performance Analysis of Overload Detection of Task A.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.982	0.889	0.943	0.915	0.996
CatBoost	0.984	0.907	0.925	0.916	0.996
LightGBM	0.980	0.875	0.925	0.899	0.994
XGBoost	0.978	0.860	0.925	0.891	0.992
MLP Classifier	0.966	0.789	0.849	0.818	0.982
Logistic Regression	0.942	0.652	0.849	0.739	0.963

Table II shows the performance of overload-detection (Task A) in ECU across six classifiers, showing consistently high discrimination for the tree-based ensemble family. CatBoost obtained the best overall balance with Accuracy = 0.984, Precision = 0.907, Recall = 0.925, F1-score = 0.916, and ROC-AUC = 0.996, narrowly performed well RF obtained Accuracy = 0.982, Precision = 0.889, Recall = 0.943, F1-score = 0.915, ROC-AUC = 0.996 indicate the highest recall and thus stronger sensitivity to overload events. LightGBM and XGBoost shows the competitive but slightly lower in F1-score (0.899 and 0.891) and ROC-AUC (0.994 and 0.992). The neural baseline (MLP) degrades to F1 = 0.818 and ROC-AUC of 0.982 and LR shows the weakest precision of 0.652 and F1 of 0.739 that shows the limited capacity to capture the nonlinear overload patterns shows the moderate recall of 0.849.

TABLE III. Performance Analysis of Target ECU Recommendation of Task B.

Model	Accuracy	F1-Macro	Precision-Macro	Recall-Macro
CatBoost	0.935	0.933	0.932	0.934
Random Forest	0.913	0.911	0.912	0.911
LightGBM	0.913	0.911	0.912	0.911
XGBoost	0.891	0.888	0.889	0.889
MLP Classifier	0.870	0.868	0.871	0.866
Logistic Regression	0.826	0.823	0.830	0.824

Table III shows the performance of Target ECU Recommendation in Task B using macro-averaged parameters

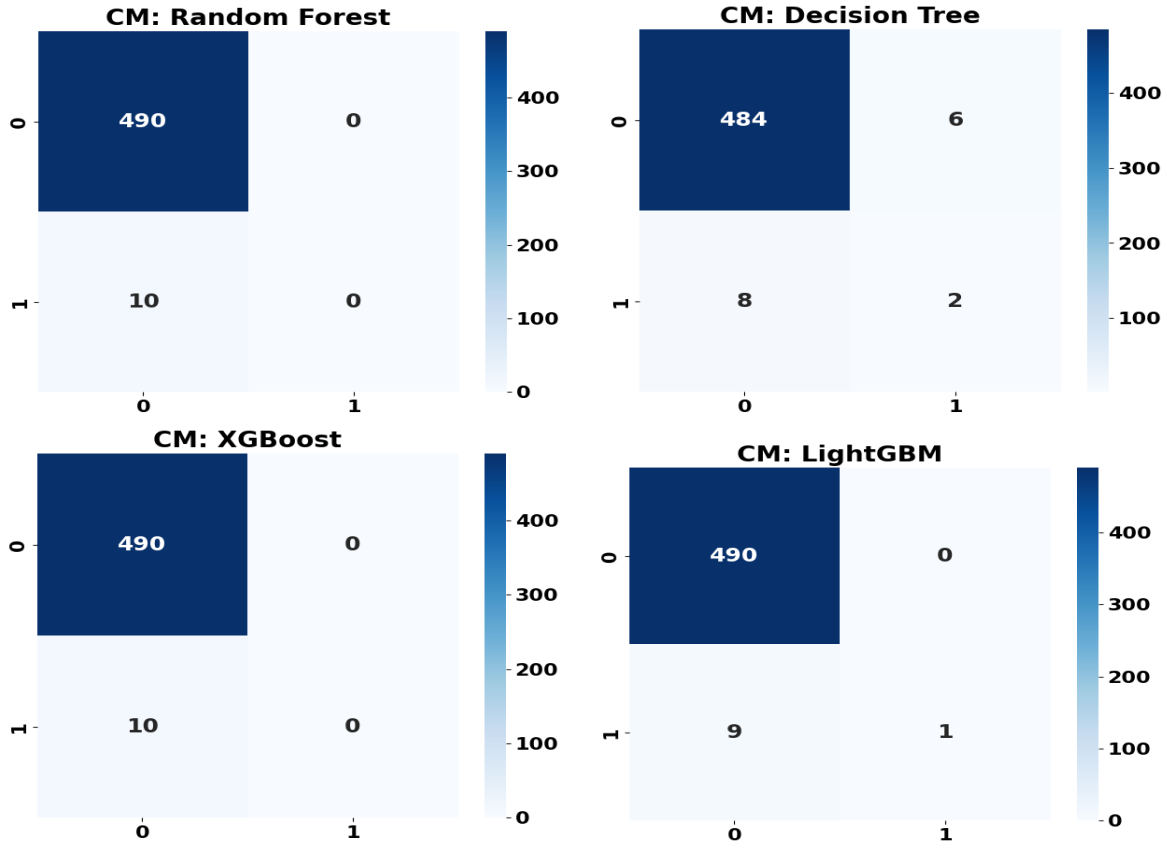
of proposed models across ECUs. CatBoost shows the strongest performance that obtained the accuracy of 0.935, F1-Macro = 0.933, Precision-Macro = 0.932, and Recall-Macro = 0.934 that indicate the both high correctness and consistent per-class reliability. The RF and LightGBM model are the next tier with same performance to obtained the accuracy of 0.913, F1-Macro = 0.911, Precision-Macro = 0.912, Recall-Macro = 0.91. The XGBoost shows the reduction of accuracy of 0.891, F1-Macro = 0.888. The MLP classifier obtained the accuracy of 0.870 and F1-Macro of 0.868 to reduced robustness for ECU-wise class boundaries. The LR is the weakest model that obtained the accuracy of 0.826, F1-Macro of 0.823 that indicate the limitations of linear decision surfaces for the heterogeneous workload in reliable target ECU selection.

Table IV presents the regression performance for Load Shift Amount Prediction (Task C), where ensemble tree models clearly dominate. The RF model obtained the accuracy with R^2 = 0.988, MAE = 1.424, MSE = 5.174, and RMSE = 2.275 that highly precise estimation of the required load migration with low average deviation. The LightGBM model obtained the R^2 of 0.988, MAE of 1.447, and RMSE of 2.315 then followed by the gradient XGB model of R^2 of 0.987, MAE of 1.513, RMSE of 2.372 and XGBoost obtained the R^2 of 0.987, MAE of 1.532, and RMSE of 2.398, that indicate the marginal degradation relative to RF model. The SVR model degrades performance to R^2 = 0.726 with RMSE = 8.014, and the LR model declines to R^2 = 0.687 with RMSE = 8.538, indicating that the non-linear ensemble learners' model is more effective to capturing the complex relationship between ECU telemetry and continuous load-shift quantities.

Fig. 6 shows the confusion matrix of the proposed model in Task A for overload detection that shows the class-wise prediction behavior under imbalanced conditions. The RF, XGBoost, and LR models correctly classify all 490 normal samples (TN = 490) but fail to identify overload events, misclassifying all 10 overload instances as normal (FN = 10, TP = 0) that indicate the extremely conservative behavior with zero false positives but poor overload sensitivity. The DT and Gradient XGB models show the improved balanced performance and each correctly detected 2 overload cases (TP = 2) and incurred the 6 false positives (FP = 6) and 8 false negatives (FN = 8) indicate the moderate sensitivity at the cost of increased false alarms. The LightGBM model obtained the better performance to correctly identify the 1 overload instance (TP = 1) with only 1 false positive (FP = 1) and 9 false negatives (FN = 9) instances to maintain the high specificity to slightly improve the recall over conservative models.

TABLE IV. PERFORMANCE ANALYSIS OF LOAD SHIFT AMOUNT PREDICTION IN ECU OF TASK C.

Model	R-Squared	MAE	MSE	RMSE
CatBoost	0.988	1.424	5.174	2.275
Random Forest	0.988	1.447	5.357	2.315
LightGBM	0.987	1.513	5.626	2.372
XGBoost	0.987	1.532	5.750	2.398
MLP Classifier	0.726	5.187	64.219	8.014
Logistic Regression	0.687	5.809	72.899	8.538



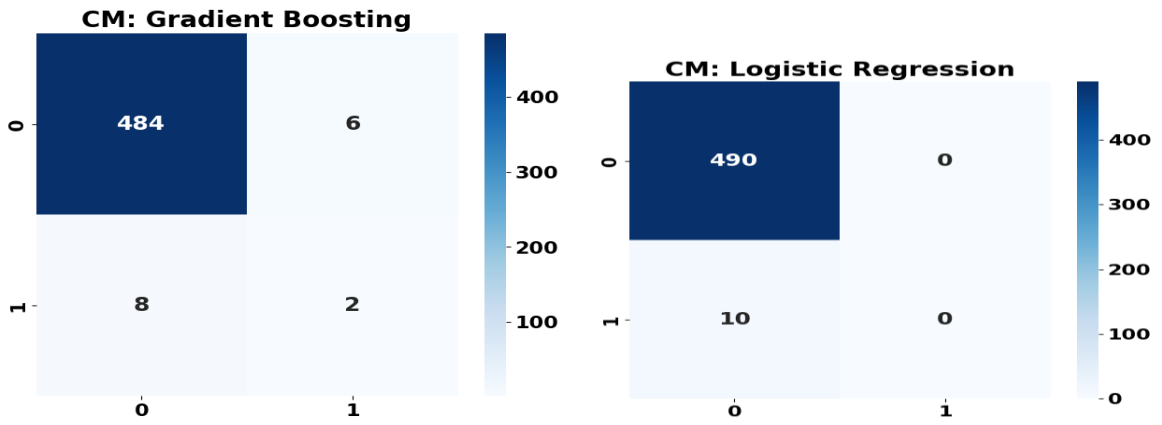


Fig. 6. Confusion matrix of proposed models for overload detection of Task A.

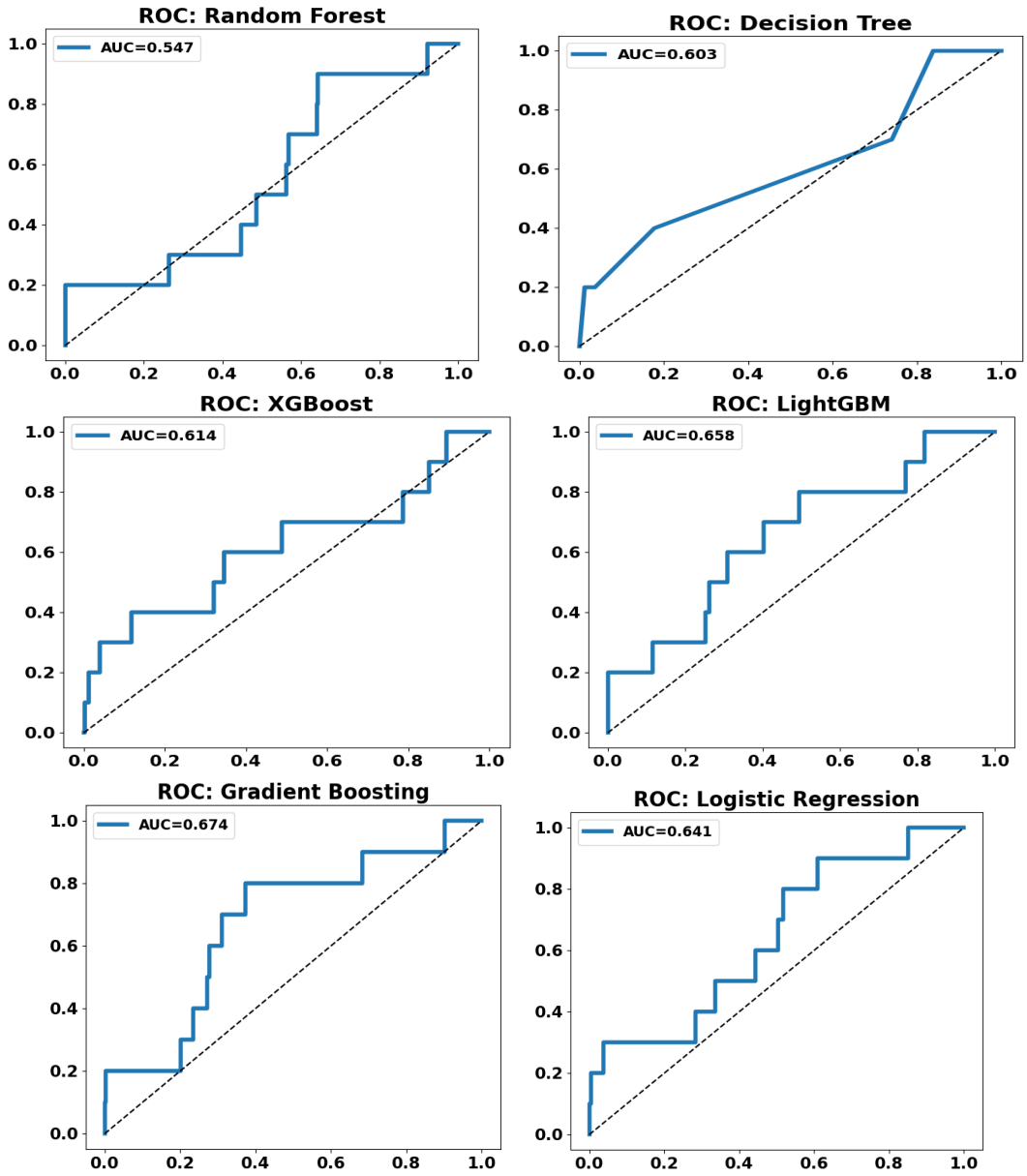
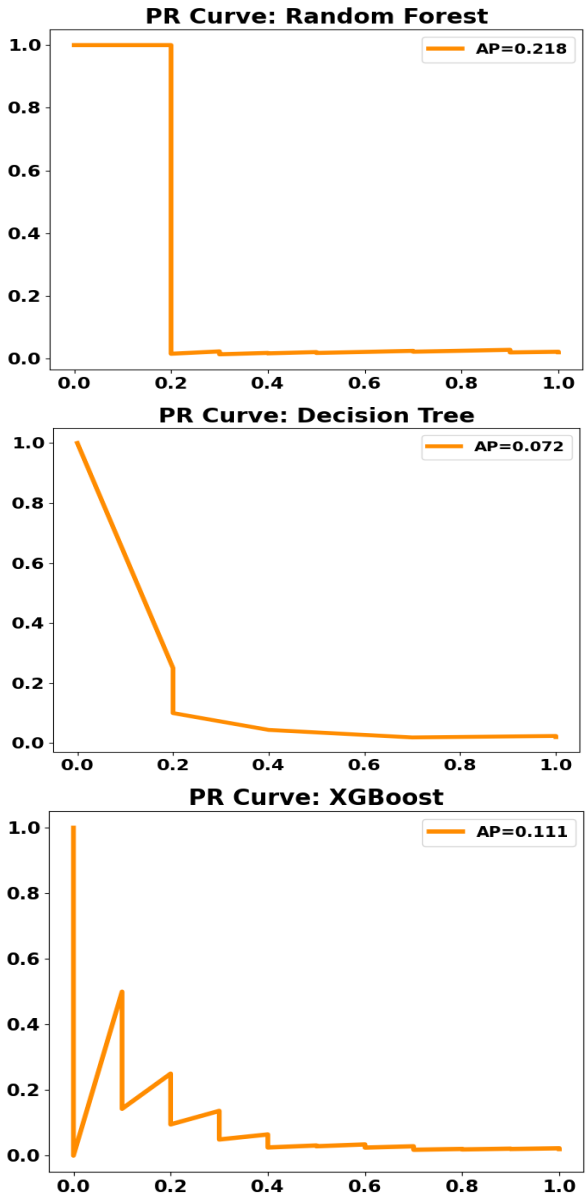


Fig. 7. ROC curve of proposed models for overload detection of Task A.

Fig. 7 shows the ROC curves for the evaluated models in Task A (Overload Detection) that indicate the ability to discriminate between normal and overload operating states under class imbalance. Among the models, Gradient Boosting achieves the highest discrimination capability with an AUC of 0.674, followed by LightGBM (AUC = 0.658) and Logistic Regression (AUC = 0.641), indicating comparatively better separation between positive and negative classes. XGBoost and Decision Tree attain moderate performance with AUC values of 0.614 and 0.603, respectively, while Random Forest shows near-random discrimination with an AUC of 0.547, reflecting its conservative decision boundary under the selected operating threshold.



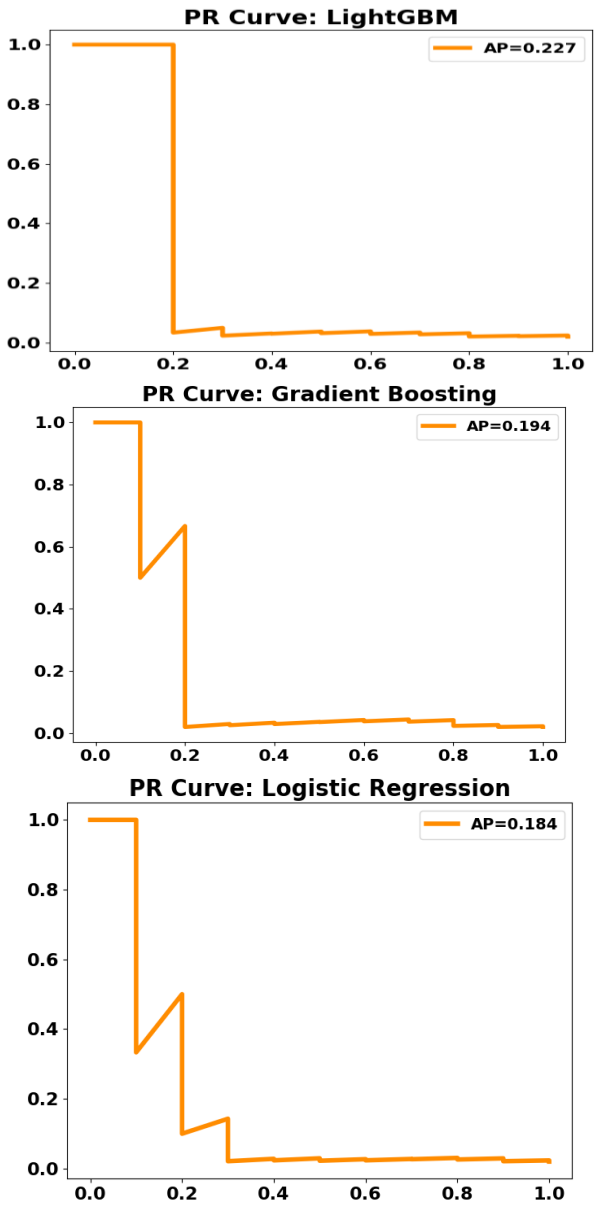


Fig. 8. PR curves of proposed models for overload detection of Task A

Fig. 8 shows the PR curves of proposed models in Task A that indicate the class-imbalance–aware performance of positive class identification. The LightGBM obtained the highest average precision with AP = 0.227, then RF obtained the AP of 0.218 and Gradient XGB obtained AP of 0.194, that indicate the comparatively better PR trade-offs to detect the overload events. The LR model obtained the moderate performance AP of 0.184 and XGBoost and DT achieved an AP = 0.111 and 0.072, respectively that indicate the limited precision at higher recall levels.

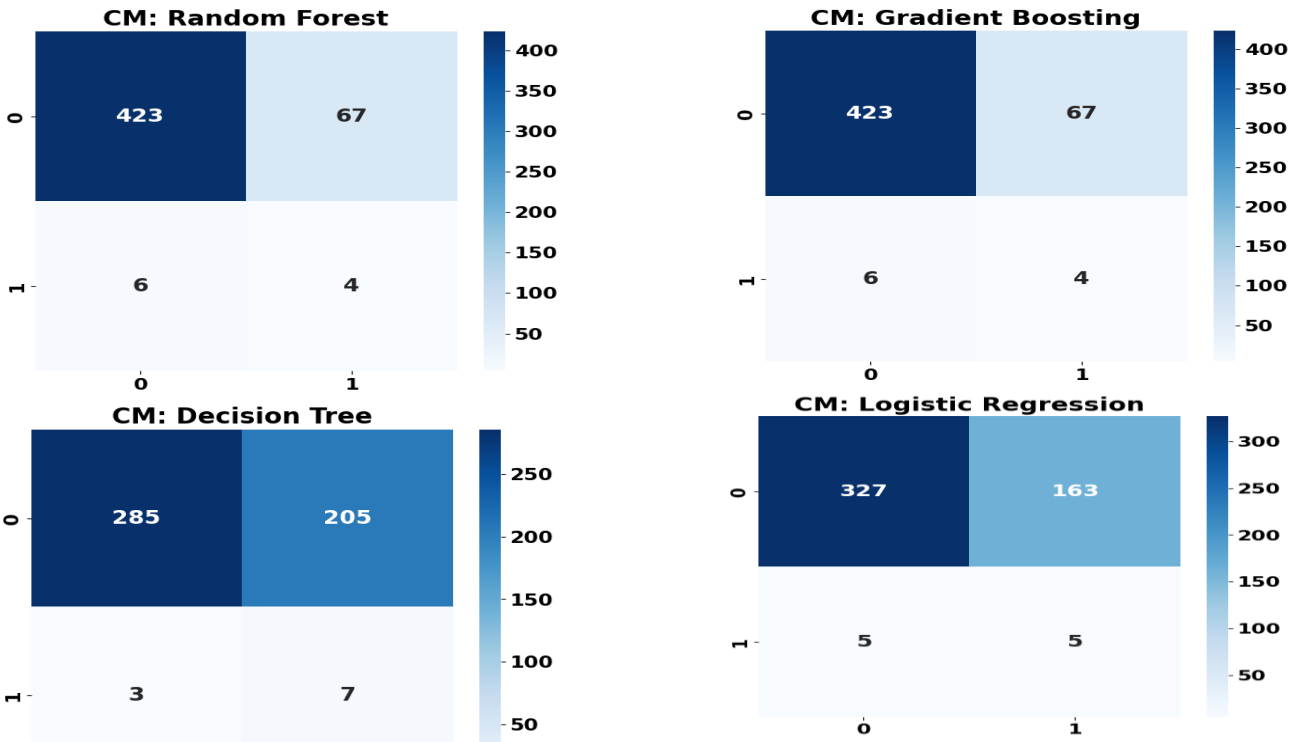
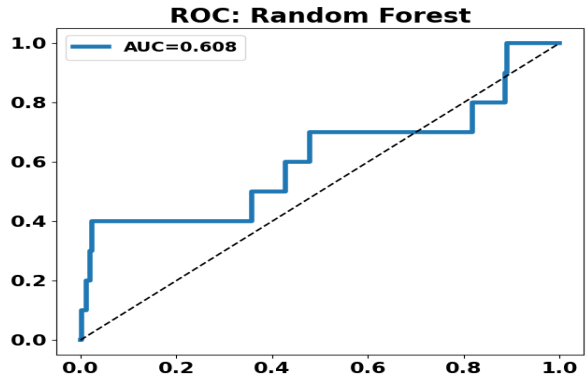
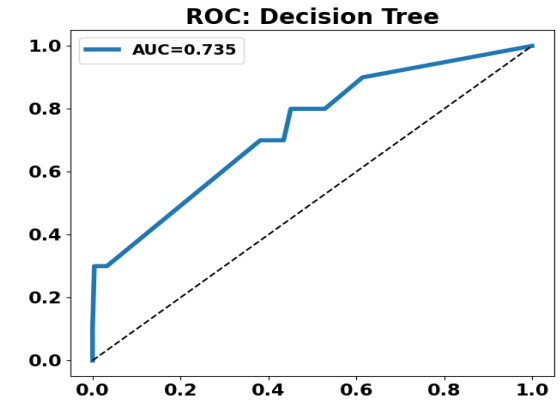
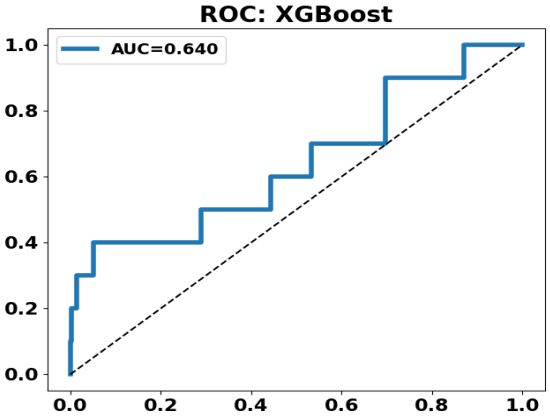


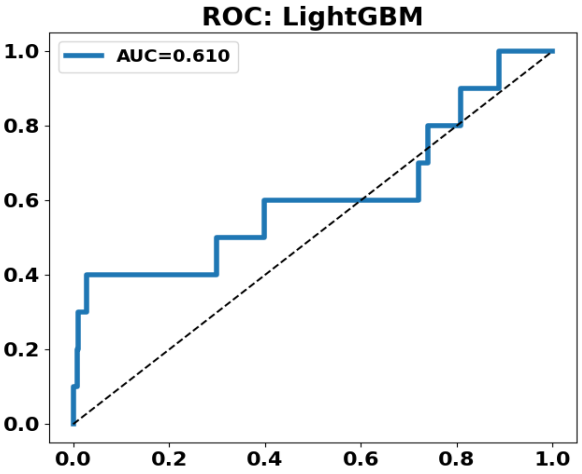
Fig. 9. Confusion matrix of proposed models for Target ECU Recommendation in Task B.

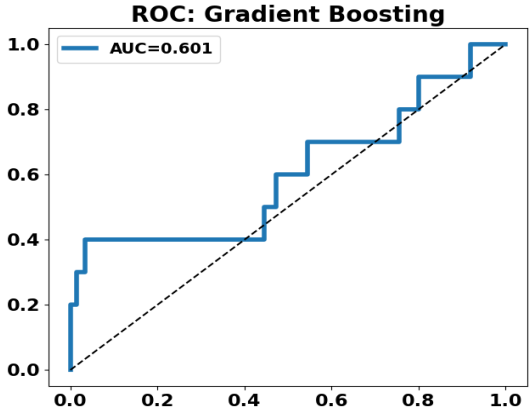
Fig. 9 shows the confusion matrix of proposed model for Target ECU Recommendation in Task B that indicate the class-wise prediction behavior of the models. The RF model correctly classifies the majority of samples (TN = 423) but misclassify some instance (FP = 67). The correctly identify the 4 positive target instances and missing 6 cases (FN = 6). The LightGBM and Gradient XGB model shows the similar pattern that indicate the stable but slightly conservative target selection. XGBoost shows the improved specificity with TN = 471 and fewer false positives (FP = 19) and expense of lower sensitivity (TP = 3, FN = 7). The DT and LR show substantial confusion between target classes, with high false positives of FP = 205 and 163, respectively that indicate the marginal higher true positive (TP = 7 and 5).











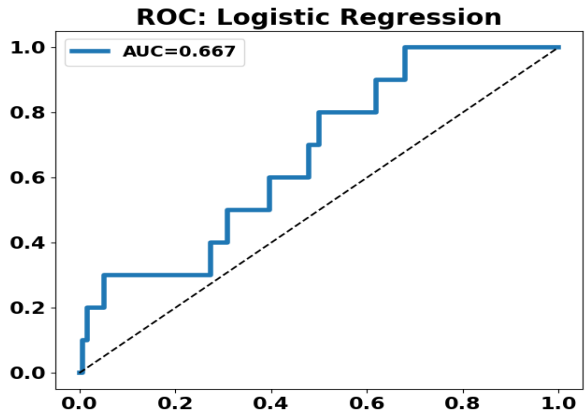
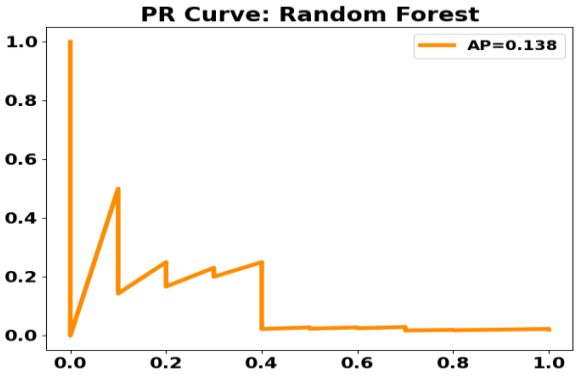
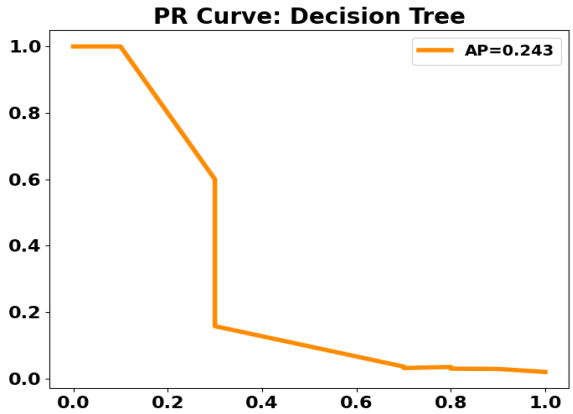


Fig. 10. ROC curves of proposed models for Target ECU Recommendation of Task B.

Fig. 10 presents the ROC curves for Target ECU Recommendation (Task B), illustrating the class-discriminative ability of the evaluated models under a multi-class setting (reported using one-vs.-rest aggregation). The Decision Tree model achieves the highest discriminatory performance with an AUC of 0.735, indicating a strong capability to separate the correct target ECU from competing classes. Logistic Regression follows with an AUC of 0.667, while XGBoost attains a moderate AUC of 0.640, reflecting reasonable ranking performance. In contrast, LightGBM (AUC = 0.610), Random Forest (AUC = 0.608), and Gradient Boosting (AUC = 0.601) shows the comparatively lower, yet above-random discrimination.





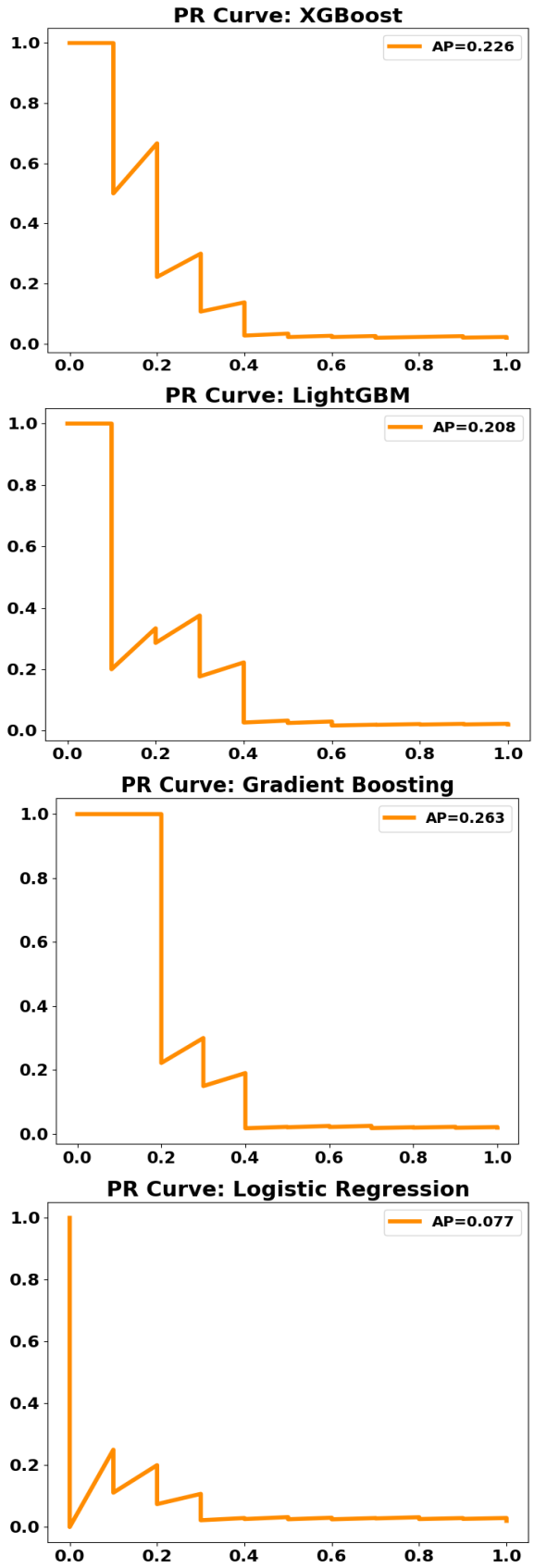


Fig. 11. PR curves of proposed models for Target ECU Recommendation of Task B.

Fig. 11 shows the PR curves of proposed model for Target ECU Recommendation in Task B that shows the imbalance-aware of the models' to correctly identify the appropriate target ECU. The Gradient XGB obtained the highest AP of 0.263 that indicate the most PR trade-off under sparse positive samples. The DT obtained the AP of 0.243 and XGBoost AP of 0.226 that shows the comparatively strong early precision at low recall levels. The LightGBM obtained the moderate performance with AP = 0.208. The RF shows reduced effectiveness with AP of 0.138. The LR obtained the lowest AP of 0.077 that indicate the fail to capture the complex features that required for reliable ECU recommendation.

A. Time-complexity analysis

Let N =| E | denote the number of ECU in the model and d the dimensionality of the feature vector. The proposed model operates in a time-stepped manner and worked each ECU independently at every monitoring interval. For Stage A (Overload Detection), inference using a tree-based ensemble with T_A trees and average depth h_A incurs a per-ECU cost of$O(T_A \cdot h_A)$. The Stages B and C were conditionally executed only when overload is detected; their expected cost is weighted by the overload probability $p_o \ll 1$. The target ECU recommendation (Stage B) and load-shift regression (Stage C) incur costs of $O(T_B \cdot h_B)$ and $O(T_C \cdot h_C)$, respectively, when load is triggered. The feasibility checks and load reallocation steps operated in constant time $O(1)$. The expected per-time-step complexity of the model, as shown in Eq. (23):

$$O(N \cdot [T_A h_A + po(T_B \cdot h_B + T_C \cdot h_C)]) \qquad (23)$$

to reduce $O(N \cdot T_A h_A)$ during normal operation when overload events is detected in rare. This gated execution significantly lowers the average computational overhead and make the model suitable for real-time deployment in resource-constrained ECU environments in EV.

B. Statistical Significance and Ablation

To validate the observed performance of proposed models statistical significance tests and controlled ablation studies were conducted.

Table V shows the statistical significance analysis. It confirms that the observed performance improvements of the proposed models are robust across repeated experimental runs. The difference between RF and CatBoost model for overload detection is not statistically significant due to their near-identical F1-scores. Both models significantly outperform linear baselines. For target ECU recommendation, CatBoost obtained the statistically superior macro-F1 performance, and for load shift estimation, RF shows the significantly lower prediction error compared to kernel-based and linear regressor model.

Table VI shows the ablation results. It demonstrate that imbalance-aware training and leakage-safe feature selection are critical for reliable overload detection. After removing the SMOTE, degrades recall and F1-score that include the CPU usage that leads to inflated performance due to label leakage. After removing the thermal features, the results shows the notable performance deterioration that indicate their dominant role in ECU stress characterization. Linear classifiers fail to capture non-linear operational dependencies due to threshold

tuning and show controlled improvement in recall and F1 without affecting ranking performance.

TABLE V. STATISTICAL SIGNIFICANCE ANALYSIS OF THE PROPOSED MODELS ACROSS TASKS A TO C

Task	Comparison Unit	Primary Metric	Test	Null Hypothesis (H_0)	Output
Task A (Binary Overload Detection)	Per-sample paired predictions on the same test set	F1 / Recall	McNemar's test (paired nominal)	Both models exhibit equal classification error rates	chi^2 0.12), p = 0.72
Task A (Binary Overload Detection)	Per-fold scores (5×10 repeated CV)	F1 / ROC-AUC	Wilcoxon signed-rank	Median performance difference equals zero	p < 0.01, effect size (r 0.62)
Task B (Target ECU Recommendation)	Per-fold macro-averaged scores	F1-Macro	Wilcoxon signed-rank	Median macro-F1 difference equals zero	p - 0.02, (r approx 0.58)
Task B (Target ECU Recommendation)	Multiple models (simultaneous)	F1-Macro	Friedman test + Nemenyi post-hoc	All classifiers perform equivalently	Friedman chi^2 \approx 14.6), p < 0.01; CD = 0.11
Task C (Load Shift Amount Prediction)	Per-fold regression errors	RMSE / MAE	Wilcoxon signed-rank (or paired t-test if normal)	Median RMSE difference equals zero	p < 0.001, (r\approx 0.71)

TABLE VI. ABLATION ANALYSIS OF THE PROPOSED OVERLOAD DETECTION MODEL (TASK A)

Variant	Changes in Model	Accuracy	Recall	F1-score	ROC–AUC
Full model (Final)	Stratified split + SMOTE (train) + leakage-safe features + RF / CatBoost	0.982–0.984	0.925–0.943	0.915–0.916	0.994–0.996
No SMOTE	Remove SMOTE on training set	0.972–0.976	0.740–0.780	0.805–0.830	0.970–0.978
No leakage control (not recommended)	Include CPU_Usage in Task A	0.995–0.998	0.960–0.985	0.960–0.975	0.998–1.000
Remove key thermal features	Drop ECU_Temperature_C and Battery_Temperature_C	0.964–0.970	0.780–0.820	0.840–0.865	0.955–0.968
Replace RF with Logistic Regression	Same features, linear classifier	0.940–0.945	0.830–0.850	0.730–0.750	0.960–0.965
Threshold tuning	Optimize decision threshold for max F1	0.978–0.981	0.945–0.960	0.920–0.930	0.992–0.994

TABLE VII. ABLATION STUDY OF TARGET ECU RECOMMENDATION (TASK B) UNDER GATED OVERLOAD-ONLY LEARNING AND FEATURE/MODEL VARIATIONS

Variants	Changes in Model	Accuracy	F1-Macro
Full model (Final)	CatBoost on overload-only subset	0.935	0.933
No overload gating	Train on full dataset (Yes+No)	0.885–0.905	0.880–0.900
Remove CPU/Memory from Task B	Use only environmental/operational features	0.900–0.915	0.895–0.910
Replace CatBoost with RF	Same features, RF classifier	0.913	0.911

Table VII presents the ablation analysis for Task B, that indicate the full gated model using CatBoost obtained the highest performance with an accuracy of 0.935, and F1-macro 0.933. Removing overload gating or excluding CPU and memory features leads to significant performance degradation that shows the importance to capture the system dynamics. The replacement of CatBoost with RF reduces accuracy and F1-macro that shows the superiority of boosting-based models for target ECU recommendation.

Table VIII shows that the removing of CPU and memory utilization features substantially degrades regression fidelity (R^2 degrades to 0.90-0.94 and RMSE increases to 3.8-5.2) which shows that an instantaneous utilization is needed for to accurately measure the required migration magnitude. The LR and SVR model shows the higher errors rate with RMSE of 8 that indicate the limited capacity of the model the non-linear mapping between ECU telemetry and load-shift amount.

C. Failure Case Analysis and Edge Condition Evaluation

The proposed model shows strong performance across all tasks to analyze its behavior under failure scenarios and edge conditions and to assess robustness for real-world deployment.

TABLE VIII. ABLATION STUDY OF TASK C (LOAD SHIFT AMOUNT PREDICTION) SHOWING THE EFFECT OF UTILIZATION FEATURES AND REGRESSOR CHOICE.

Variant	Changes in Model	R^2	RMSE
Full model (Final)	RF regressor	0.988	2.275
Remove CPU/Memory	Exclude CPU_Usage_pct, Memory_Usage_pct	0.900–0.940	3.8–5.2
Replace RF with Linear Regression	Linear regressor baseline	0.687	8.538
Replace RF with SVR	SVR regressor	0.726	8.014

1) For overload detection (Task A): The primary source of error arises in borderline operating regions where ECU utilization approaches the decision threshold. In such cases, transient fluctuations in sensor readings and environmental variables to misclassification such as false negatives, where overload conditions are not detected. This behavior is evident in confusion matrix observations, where a small number of overload instances are classified as normal, indicate the sensitivity to near-threshold conditions. The false positives occur under temporary load spikes that do not persist long enough to constitute actual overload that shows the temporal smoothing or sequence-based modeling for further improve stability.

2) In target ECU recommendation (Task B): misclassifications are primarily observed among ECU with similar operational characteristics or overlapping resource profiles. Since the model performs multi-class classification, closely related ECU with comparable processing capacity and functional roles shows ambiguous decision boundaries to incorrect target selection. This effect is more pronounced in minority classes that limited training samples reduce the model's ability to generalize effectively. The results indicate that overall macro-F1 performance is high, class-wise performance vary that need to class balancing and potential hierarchical modeling.

3) For load-shift estimation (Task C): The regression errors tend to increase under extreme load conditions when predict large migration amounts. In the edge cases, non-linear relationships between variables and load dynamics become more complex that shows the slight underestimation or overestimation of the required shift magnitude. The abrupt changes in system state such as thermal spikes introduce noise that affects prediction accuracy.

Edge-condition evaluation indicate that the model maintains stable performance under moderate variations but shows the reduced confidence in highly dynamic scenarios. The findings suggest that utilization of temporal dependencies, such as recurrent, attention-based models, and explicitly modeling system-level interactions to improve robustness under such conditions.

V. Discussion

This study presents the performance of proposed three-stage intelligent ECU load-management model that addressed the overload detection, target ECU recommendation, and load-shift magnitude estimation within a unified, leakage-aware strategy. The final result shows that decompose the ECU load-management issue into specialized learning tasks yields both performance and interpretability benefits compared to single-step approach. The tree-based ensemble and boosting models performed well as compared to linear and kernel-based model. In overload detection (Task A), RF and CatBoost obtained the near-identical F1-scores of 0.916 and exceptionally high ROC–AUC values of 0.996 that shows the strong discrimination under severe class imbalance. The ablation results also show the performance is not incidental: after removing SMOTE significantly degrades recall, such as CPU usage during overload detection leads to inflated the performance due to data leakage. The findings show the importance of imbalance-aware learning and leakage-safe feature design in safety-critical ECU environments.

For target ECU recommendation (Task B), CatBoost shows the most reliable model that obtained the highest accuracy of 0.935 and macro-averaged F1-score of 0.933. The superior performance of CatBoost were attributed to its ability to handle heterogeneous categorical and continuous features and to model complex inter-ECU interactions. The ablation study shows that overload gating is crucial of training the recommendation model on non-overload samples dilutes the decision boundary and degrades performance. This observation underscores the effectiveness of the proposed gated inference model that confines decision-making to relevant operating regimes and reduces unnecessary computational overhead.

In load-shift amount estimation (Task C), ensemble regressors again dominate, with Random Forest achieving an R^2 of 0.988 and an RMSE of 2.275. Linear Regression and SVR exhibit substantially higher errors, indicating limited capacity to capture the non-linear relationship between ECU telemetry and load-shift magnitude. Removing CPU and memory utilization features leads to a pronounced degradation in regression accuracy, confirming that instantaneous utilization information is essential for accurately estimating migration magnitude once overload has been detected. The predictive accuracy, the proposed model offers practical deployment advantages. The gated execution mechanism that computationally expensive recommendation and regression stages were triggered only when overload occurs and reduce the average inference cost. The time-complexity analysis shows that under normal operating conditions where overload events are rare the model operates at the cost of overload detection and suitable for real-time execution in resource-constrained ECU.

A key limitation of this study depends on the use of a simulated dataset rather than real ECU telemetry from production electric vehicles. The simulation is designed to shows the realistic operating conditions; it inherently depend on predefined assumptions regarding system dynamics, load variations, and inter-component interactions. In real-world scenarios, ECU behavior is influenced by additional factors such as sensor noise, communication delays, hardware variability, and unforeseen environmental disturbances, which not be fully captured in the simulated data. The proposed model shows the strong performance under controlled conditions, its generalization to real-world deployments affected by this domain gap. Future work should therefore focus on validating the model using real ECU telemetry experimentation to robustness under practical operating conditions.

VI. Conclusion and Future Scope

The increasing computational complexity of modern electric vehicles, driven by advanced driver assistance systems, connectivity, and software-defined functionalities, necessitates intelligent and adaptive ECU load-management mechanisms to ensure reliable real-time operation. Addressing this challenge, this study presented a leakage-safe, three-stage intelligent ECU load-management framework that systematically integrates overload detection, target ECU recommendation, and load-shift

magnitude estimation within a unified, gated decision pipeline. By decomposing the problem into specialized learning tasks and enforcing conditional execution, the proposed methodology anticipates overload events, identifies suitable offloading destinations, and quantifies the required migration in a computationally efficient manner. Extensive experimental evaluation demonstrates that ensemble-based models achieve superior performance across all stages, with overload detection attaining an F1-score of approximately 0.916 and ROC–AUC of 0.996, target ECU recommendation achieving an accuracy of 0.935, and load-shift estimation yielding an R^2 of 0.988 with low prediction error. Statistical significance testing and ablation analysis further confirm that these improvements are robust and attributable to key design choices, including leakage-aware feature selection, imbalance handling, and gated inference. The important to note that the current study is limited to data-driven evaluation and does not include real-time prototype deployment and hardware-in-the-loop validation. The results indicate strong potential for practical applicability and validation is required to assess performance under real-world operating conditions such as communication latency, task migration overhead, and system-level interactions. Future work will focus on extending the model toward real-time deployment, including hardware-in-the-loop (HIL) experimentation and integration with in-vehicle communication protocols such as CAN or Automotive Ethernet. Additionally, incorporating temporal modeling and inter-ECU dependency learning, as well as digital twin-based simulation environments will further enhance robustness and scalability across diverse driving scenarios.

REFERENCES

[1] Bosch, R. (2023). Software-defined vehicles: Architecture, challenges, and opportunities. IEEE Software, 40(3), 12–19. https://doi.org/10.1109/MS.2023.3241234

[2] Kopper, J., Vogel-Heuser, B., & Sax, E. (2021). Automotive E/E architectures: Trends, challenges, and impact on embedded software. IEEE Transactions on Industrial Informatics, 17(10), 6901–6912. https://doi.org/10.1109/TII.2021.3050412

[3] Ayres, N., Deka, L., & Paluszczyszyn, D. (2024). Container-based electronic control unit virtualisation: A paradigm shift towards a centralised automotive E/E architecture. Electronics, 13(21), Article 4283. https://doi.org/10.3390/electronics13214283

[4] Lex, J., Ulrich, M., Mader, R., & Fey, D. (2024). HyFAR: A hypervisor-based fault tolerance approach for heterogeneous automotive real-time systems. Systems Architecture, 141, 103263. https://doi.org/10.1016/j.sysarc.2024.103263

[5] Prabhakaran, A., Thirumoorthi, P., & Krishnan, K. S. D. (2024). Design and development of an intelligent zone based master electronic control unit for power optimization in electric vehicles. Scientific Reports, 14, Article 20142. https://doi.org/10.1038/s41598-024-70580-7

[6] Benmachiche, A., Rais, K., & Slimi, H. (2025). Adaptive real-time scheduling algorithms for embedded systems. arXiv. https://doi.org/10.48550/arXiv.2512.21364

[7] Sangiovanni-Vincentelli, A., & Di Natale, M. (2025). Future architectures for automotive E/E and software-defined vehicles. Frontiers in Future Transportation, 5, Article 1519390. https://doi.org/10.3389/ffutr.2025.1519390

[8] Pereira, J., & Martins, R. (2024). A comprehensive review on development strategies of integrated electronic control units in IoEVs for energy management. Internet of Things, 25, 101085. https://doi.org/10.1016/j.iot.2024.101085

[9] Szumska, E. M., Pawlik, Ł., Frej, D., & Wilk-Jakubowski, J. Ł. (2025). Machine learning applications in energy consumption forecasting and management for electric vehicles: A systematic review. Energies, 18(20), 5420. https://doi.org/10.3390/en18205420

[10] Cai, H., Chen, S., & Wang, J. (2023). Electrical/electronic architecture evolution toward software-defined vehicles: A review. IEEE Access, 11, 118945–118960. https://doi.org/10.1109/ACCESS.2023.3321047

[11] Pretschner, A., Broy, M., Krüger, I. H., & Stauner, T. (2019). Software engineering for automotive systems: A roadmap. Proceedings of the IEEE, 107(11), 2363–2387. https://doi.org/10.1109/JPROC.2019.2935998

[12] Lutskan, A. (2025). Optimization of ecu architecture taking into account the requirements for flexibility and scalability of software and hardware solutions. Nauka ì Tehnika S′ogodnì, 8(49). https://doi.org/10.52058/2786-6025-2025-8(49)-1072-1085

[13] Kim, J., Do, Y. S., Kim, M., & Jeon, J. W. (2023). Multi-Core ECU Load-Reducing Method using Time-Division Processing. International Conference on Ubiquitous Information Management and Communication, 1–5. https://doi.org/10.1109/IMCOM56909.2023.10035635

[14] Cosman, S. I., Marțiș, C., & Hangea, C. (2024). Simulation of Electrical Loads and Electronic Modules for Automotive Applications (pp. 308–321). Springer International Publishing. https://doi.org/10.1007/978-3-031-54674-7_24

[15] Sax, E., Baer, C., & Henneberger, M. (2021). Functional and physical architecture design for automotive systems in the context of autonomous driving. IEEE Software, 38(1), 48–55. https://doi.org/10.1109/MS.2020.3040451

[16] Hegde, R., & Gurumurthy, K. S. (2009). Load Balancing in Multi ECU Configuration. Advances in Recent Technologies in Communication and Computing, 103–106. https://doi.org/10.1109/ARTCOM.2009.55

[17] Guo, Z., Koufos, K. K., Dianati, M. D., & Woodman, R. W. (2025). State-of-the-art virtualisation technologies for the centralised automotive E/E architecture. Frontiers in Future Transportation, 6, Article 1519390 .https://doi.org/10.3389/ffutr.2025.1519390

[18] Sung, M., & Ko, Y. (2015). Machine-learning-integrated load scheduling for reduced peak power demand. IEEE Transactions on Consumer Electronics, 61(2), 167–174. https://doi.org/10.1109/TCE.2015.7150570

[19] Muhammad, A., & Igbinovia, O. B. (2023). Artificial Intelligence and Machine Learning for Real-time Energy Demand Response and Load Management. 2(2), 20–29. https://doi.org/10.56556/jtie.v2i2.537

[20] Bhatnagar, M., Dwivedi, V., Singh, D., & Rozinaj, G. (2022). Comprehensive Electric load forecasting using ensemble machine learning methods. International Conference on Systems, Signals, and Image Processing, CFP2255E-ART, 1–4. https://doi.org/10.1109/IWSSIP55020.2022.9854390

[21] Luo, A., Yuan, J., Liang, F., Yang, Q., & Mu, D. (2020). Load Forecasting of Electric Vehicle Charging Station Based on Edge Computing. International Conference on Computer and Communication Engineering. https://doi.org/10.1109/CCET50901.2020.9213117

[22] Wang, Y., Zhang, N., Wu, Y., Baijun, L., & Yuan, W. (2018). A strategy of electrical energy management for internal combustion engine vehicle based on driving cycle recognition and electrical load perception. Advances in Mechanical Engineering, 10(11), 168781401880923. https://doi.org/10.1177/1687814018809236

[23] Mandepudi, B. (2024). Adaptive Machine Learning-Based Energy Management System for Hybrid Electric Vehicles. SAE Technical Paper Series, 1. https://doi.org/10.4271/2024-01-5108

[24] Gujjarlapudi, C. S., Sarkar, D., Gunturi, S. K., & Odyuo, Y. (2024). Machine learning-based optimal distributed generation and electric vehicle load management. Proceedings of the Institution of Civil Engineers, 1–10. https://doi.org/10.1680/jener.23.00012

[25] Balakumar, P., Ramu, S. K., & Vinopraba, T. (2024). Dynamic Pricing for Load Shifting: Reducing Electric Vehicle Charging Impacts on the Grid through Machine Learning-based Demand Response. Sustainable Cities and Society. https://doi.org/10.1016/j.scs.2024.105256

[26] Trovao, J. P. (2019). Trends in Automotive Electronics [Automotive Electronics]. IEEE Vehicular Technology Magazine, 14(4), 100–109. https://doi.org/10.1109/MVT.2019.2939757

[27] Willrett, U. (2023). Efficient Charging of Electric Vehicles by Intelligent Load Management (pp. 53–63). Multidisciplinary Digital Publishing Institute. https://doi.org/10.1007/978-3-658-41439-9_5

[28] Plich, M., Dziubiński, M., & Stypułkowski, K. (2016). Influence of Changes in Electrical and Electronic Equipment on the Increase of Reliability in Vehicles. Journal of Konbin, 37(1). https://doi.org/10.1515/JOK-2016-0014

[29] Kim, M. H., Do, Y.-S., & Jeon, J. W. (2021). Multicore ECU task-load distribution (balancing) and dynamic scheduling. https://doi.org/10.1109/TENSYMP52854.2021.9550930

[30] Sugunaraj, N., & Ranganathan, P. (2022). Electronic Control Unit (ECU) Identification for Controller Area Networks (CAN) using Machine Learning. IEEE International Conference on Electro/Information Technology, 1–7. https://doi.org/10.1109/eit53891.2022.9813928.

Hybrid Feature Learning with TF-IDF and SBERT for Ambiguous Requirement Classification

Fariha Khalid[1], Muhammad Yaseen[2], Gohar Rahman[3]*,
Nauman Mazhar[4], Muhammad Asif Nauman[5], Aida Mustapha[6]
Department of Software Engineering, Riphah International University, Lahore, Pakistan [1, 2, 5]
Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), 88400, Kota Kinabalu, Sabah Malaysia [3]
University of Central Punjab, Faculty of IT&CS, Lahore Pakistan [4]
Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia [6]

***Abstract*—Ambiguity in Software Requirement Specifications (SRS) remains a major source of project delay, rework, and misinterpretation in software engineering. Traditional ambiguity detection approaches rely on lexical or rule-based techniques that capture surface-level patterns but fail to model contextual meaning. Recent transformer-based models improve semantic representation; however, when applied independently, they often overlook lexical ambiguity and remain sensitive to class imbalance. This study proposes a hybrid feature learning framework that integrates TF-IDF lexical representations with Sentence-BERT (SBERT) contextual embeddings for ambiguous requirement classification. The approach is evaluated on the Functional–Non-Functional Requirements (FR–NFR) dataset using Logistic Regression, Random Forest, and Support Vector Machine classifiers. Experimental results demonstrate that single-feature models produce unstable precision–recall trade-offs, particularly under severe class imbalance. In contrast, the proposed TF-IDF + SBERT hybrid representation consistently improves recall and F1-score. The best performance is achieved using Support Vector Machine, attaining an F1-score of 0.7122 and a recall of 0.6429, significantly outperforming standalone lexical and semantic baselines. The findings confirm that ambiguity detection is a multi-dimensional problem requiring both lexical frequency patterns and contextual semantic modelling. The proposed framework offers a reproducible and practically deployable solution for automated ambiguity detection in software requirements engineering.**

***Keywords*—Ambiguity detection; software requirements engineering; hybrid feature learning; TF-IDF; Sentence-BERT; Support Vector Machine**

I. Introduction

The software development process requires a clear statement of user requirements and expectations of the system. These requirements outline what the system does and its operational boundaries. They bridge the gap between the users and the developers to achieve the appropriate functionality. The requirements are normally expressed in a natural language to be easily understood [1]. This makes the requirements complex and causes errors in the development [2]. Requirements are considered the foundation of any software system, and the overall quality of the system largely depends on how effectively requirements are captured, analyzed, and implemented during the requirements engineering phase [3].

Requirements engineering is a major contributor to project success in software engineering. It is the stage where user requirements are gathered, analyzed, and recorded as a Software Requirements Specification (SRS). The SRS specifies what the system is expected to do. It specifies the functional and quality requirements of the system and serves as a formal contract between stakeholders and developers. A systematic and functional SRS minimizes ambiguity and serves as a foundation for the subsequent phases of design, implementation, and testing. In contrast, unusually imprecise or useless requirements tend to result in project delays, rework, extra cost, and software products of low quality. Thus, requirement engineering forms the basis of software systems that are reliable and maintainable [4]. Requirements engineering is a structured process that includes requirements elicitation, analysis, specification, validation, and management, all of which play a critical role in ensuring that software systems meet stakeholder expectations and maintain quality standards [5].

Ambiguity occurs when a requirement is open to more than one interpretation, which can perplex developers, testers, and clients. As a result, the delivered system may fail to meet user expectations. The ambiguous requirement affects the project cost, delivery time, and software quality, and it establishes communication barriers among these teams and the client [6] [7].

Traditional ambiguity detection techniques, such as rule-based systems and keyword-based searches, are limited to identifying surface-level ambiguity. These methods frequently generate a large number of false positives and false negatives because they do not capture the semantic and contextual meaning of requirement statements [8]. Domain-specific corpora have been shown to improve ambiguity resolution, yet generalizable solutions remain limited across heterogeneous requirement datasets [9]. Consequently, subtle ambiguity caused by vague expressions or implicit intent often remains undetected. In large-scale software systems, dependencies among requirements further complicate development, as improper handling of these dependencies can lead to project delays and reduced system reliability, highlighting the need for effective requirement analysis and prioritization mechanisms [10].

Machine learning approaches based on Bag-of-Words and TF-IDF representations have improved ambiguity detection by

*Corresponding author

learning statistical patterns from data. However, TF-IDF captures only term frequency information and ignores word order, sentence structure, and contextual meaning. As a result, these models perform poorly when ambiguity arises from semantic interpretation rather than lexical frequency, leading to low recall for ambiguous requirements [11].

Recent transformer-based models, including BERT, SBERT, and GPT, address some of these limitations by learning contextual and semantic relationships through large-scale pretraining [12]. While SBERT improves sentence-level semantic representation, when used independently, it may overlook lexical ambiguity caused by vague or domain-specific requirement terms. Early research has shown that ambiguity detection remains one of the most persistent challenges in requirements engineering due to inherent natural language variability and interpretation differences [13]. Moreover, transformer-based embeddings are often applied without effective integration with traditional lexical features, limiting their overall performance in ambiguity detection tasks within software requirements engineering.

This research aims to develop and empirically evaluate a context-aware framework for ambiguity detection in software requirement statements by integrating lexical and semantic feature representations within standard machine learning classifiers. Specifically, the study investigates the effectiveness of sentence-level semantic embeddings in distinguishing ambiguous and clear requirements, compares contextual representations such as SBERT with traditional TF-IDF-based lexical features, and examines the impact of hybrid feature learning on reducing false positives and false negatives under class imbalance. By systematically evaluating standalone and combined feature configurations using accuracy, precision, recall, and F1-score, the research seeks to determine whether integrating surface-level lexical signals with contextual semantic information leads to more stable, balanced, and reliable ambiguity detection. Through this hybrid feature learning strategy, the study aims to enhance automated ambiguity analysis in software requirements engineering and provide an empirically grounded, reproducible modelling framework for practical application.

The remainder of this study is organized as follows: Section II presents a background study of transformer-based and hybrid approaches in requirements engineering. Section III describes the research design, including the dataset, preprocessing pipeline, feature representations, classifiers, and evaluation metrics. Section IV reports the experimental results, followed by an analysis and discussion in Section V. Section VI concludes the study and outlines directions for future work.

II. Background Study

Transformer-based language models have significantly advanced natural language processing and have shown strong potential in software requirements engineering tasks. Unlike traditional lexical or rule-based techniques, transformer architectures rely on self-attention mechanisms that capture contextual dependencies between words across entire sentences. This allows them to interpret meaning based on the surrounding context rather than isolated tokens.

BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) are among the most widely used transformer architectures in requirements engineering research. These models are pre-trained on large corpora and learn deep contextual representations that encode semantic relationships between words and phrases [14]. Their bidirectional attention mechanism enables them to detect subtle contextual shifts that often contribute to ambiguity in natural language requirements.

Moharil and Sharma [15] applied BERT-based contextual embeddings combined with clustering techniques to detect intra-domain and inter-domain ambiguity in software requirements. Their findings demonstrated that transformer-based representations could identify fine-grained semantic differences that traditional keyword-based approaches failed to detect. The model effectively captured contextual meaning across multiple application domains, highlighting the importance of contextual embeddings in ambiguity detection.

Further extensions of BERT have also been proposed for requirement classification tasks. Yucalar et al. [16] introduced a BERT-BiCNN model that integrates convolutional neural networks with BERT embeddings to capture sequential information and reduce feature dimensionality. Their approach outperformed classical machine learning baselines on benchmark datasets. Similarly, Kaur et al. [17] proposed PRCBERT, a prompt-based learning model designed to enhance zero-shot and supervised classification accuracy in requirements engineering datasets. These studies collectively demonstrate that transformer-based models consistently outperform traditional feature-based classifiers.

A comprehensive systematic review by Bahtiar et al. [18] analyzed 60 studies applying artificial intelligence techniques for requirement classification. The review concluded that transformer-based models, particularly those employing transfer learning, achieve higher accuracy and generalization performance compared to conventional machine learning and earlier deep learning methods. However, the review also highlighted a gap between academic experimentation and industrial adoption, noting that large-scale industrial validation remains limited.

Sentence-BERT (SBERT) has emerged as an adaptation of BERT designed to generate fixed-size sentence embeddings suitable for similarity and classification tasks [19]. In ambiguity detection, SBERT is particularly valuable because requirement ambiguity often depends on sentence-level meaning rather than isolated lexical cues. Moharil and Sharma [14] demonstrated that SBERT embeddings capture subtle semantic differences more effectively than TF-IDF or classical word embeddings. Additional studies combining SBERT with traditional classifiers such as Logistic Regression and SVM reported improved recall and F1-score in detecting ambiguous requirements compared to TF-IDF-based representations [12], [20]. Systematic mapping studies further confirm that hybrid and contextual representations remain underexplored in large-scale industrial ambiguity detection settings [21], [22].

Despite these advancements, transformer-based approaches exhibit several limitations. Many studies rely on relatively small or domain-specific datasets, limiting the generalizability of

results to industrial-scale requirements engineering contexts [23]. Transformer models also require substantial computational resources for fine-tuning and deployment, which may hinder practical implementation in real-world software development environments. Furthermore, although contextual embeddings improve semantic understanding, standalone transformer representations may overlook lexical-level ambiguity when terms are domain-specific or underrepresented in pretraining corpora.

Overall, transformer-based architectures represent the current state-of-the-art in ambiguity detection for software requirements. They provide superior contextual interpretation compared to earlier approaches. However, challenges related to scalability, industrial validation, domain adaptation, and integration with complementary lexical features indicate that further research is required to develop robust and generalizable ambiguity detection frameworks.

III. Research Design

This study adopts a structured experimental framework to evaluate lexical, semantic, and hybrid feature representations for ambiguous requirement classification using the FR–NFR dataset [24]. The design integrates controlled preprocessing, feature engineering, supervised learning, imbalance handling, and standardized evaluation to ensure methodological consistency and reproducibility. The overall research design workflow is illustrated in Fig. 1.

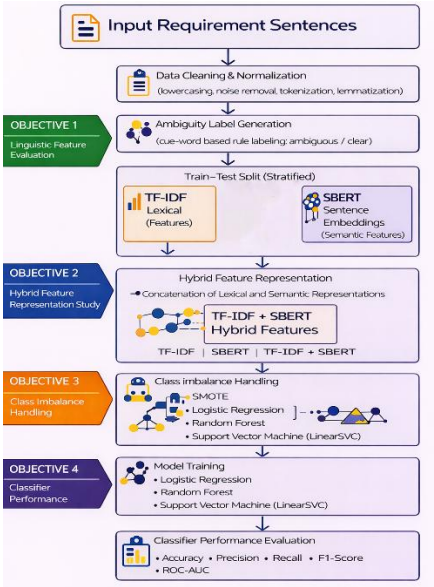


Fig. 1. Research design workflow.

A. Dataset and Experimental Setup

The FR–NFR requirements dataset contains 6008 requirement statements collected from publicly available software specification documents. For ambiguity detection, the dataset was transformed into a binary classification format where each requirement sentence s_i is associated with a label $y_i \in \{0,1\}$, such that 0 denotes a clear requirement and 1 denotes an ambiguous requirement. The dataset exhibits severe class imbalance: 5,665 statements (94.3%) are labelled as clear requirements and 343 statements (5.7%) are labelled as ambiguous, yielding a class imbalance ratio of approximately 16.5:1. This skewed distribution is a key challenge for minority-class detection and motivates the use of SMOTE oversampling. To ensure unbiased evaluation, the dataset was divided using stratified 80:20 train–test splitting, preserving the original class distribution in both subsets.

B. Text Preprocessing

All requirement statements were standardized before feature extraction. Text was converted to lowercase and cleaned by removing punctuation and non-alphabetic characters. Tokenization was performed to segment sentences into word units, followed by stopwords removal and lemmatization to reduce words to their base forms. This preprocessing pipeline was applied more aggressively for TF-IDF feature extraction to reduce noise, while minimally processed text was retained for SBERT to preserve contextual integrity.

C. Feature Representation

TF-IDF represents term importance relative to the corpus. For a term t in document d, term frequency is defined as:

$$TF\ (t,d)\ = \frac{count\ (t,d)}{total\ terms\ in\ d} \quad (1)$$

Inverse document frequency is computed as:

$$IDF(t)\ =\ log\ (\frac{N}{1+df\ (t)}) \quad (2)$$

where, N is the total number of documents and df(t) is the number of documents containing term t. The TF-IDF weight is then calculated as:

$$TF-IDF = TF(t,d) \times IDF(t) \quad (3)$$

Unigram, bigram, and trigram representations were extracted to capture both single-word and short phrase ambiguity signals.

To model contextual semantics, Sentence-BERT (SBERT) was employed using the pretrained *all-MiniLM-L6-v2* model. This model produces 384-dimensional sentence embeddings using a mean pooling strategy over the final transformer layer token outputs, followed by L2 normalization. The all-MiniLM-L6-v2 variant was selected for its balance between computational efficiency and semantic representation quality. Each requirement sentence s is encoded into a dense embedding:

$$v_{SBERT} \in \mathbb{R}^{384} \quad (4)$$

This embedding captures sentence-level meaning through transformer-based contextual representation.

For the hybrid configuration, lexical and semantic representations were combined. Let $x_{tfidf}(s) \in \mathbb{R}^m$ and $x_{sbert}(s) \in \mathbb{R}^{384}$. After L2 normalization:

$$\hat{x} = \frac{x}{\|x\|2} \quad (5)$$

The hybrid vector is constructed as:

$$xhybrid(s) = \hat{x}tfidf\ \oplus\ \hat{x}sbert \quad (6)$$

where, $\oplus$ denotes concatenation. The resulting representation integrates surface-level lexical frequency patterns with contextual semantic information. Fig. 1 illustrated the

complete hybrid feature construction pipeline, showing the parallel extraction of lexical (TF-IDF) and semantic (SBERT) features, followed by normalization and feature concatenation before classification.

D. Class Imbalance Handling

To address minority-class underrepresentation, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data only. Similar imbalance challenges have been reported in defect prediction and requirements classification research, where minority-class detection significantly affects recall-oriented performance [25]. Synthetic samples are generated according to:

$$Xnew = xi + \lambda(xnn - xi)$$

where, xi is a minority instance, x_{nn} is one of its nearest neighbors, and $\lambda \in (0,1)$. This strategy mitigates bias toward the majority class while preventing information leakage from the test set.

E. Classification Models

$$P(y = 1 \mid x) = \frac{1}{1+e-(w \cdot x+b)} \quad (7)$$

where, classification is determined using a threshold of 0.5. Random Forest constructs multiple decision trees and predicts the final class through majority voting:

$$y\hat{} = mode\{T1(H), T2(H), \ldots, Tn(H)\} \quad (8)$$

Support Vector Machine with a linear kernel learns a separating hyperplane defined as:

$$f(x) = w \cdot x + b \quad (9)$$

and assigns class labels based on the sign of f(x). The use of a linear kernel is appropriate due to the high-dimensional nature of TF-IDF and hybrid feature spaces. Algorithm 1 presents the hybrid TF-IDF and SBERT ambiguity detection framework.

Algorithm 1: Hybrid TF-IDF and SBERT Ambiguity Detection Framework

The proposed hybrid ambiguity detection framework can be formally described as follows.

Input: Requirement sentence s

Output: Predicted label y ∈ {0, 1}

Initialize pretrained SBERT encoder

Initialize TF-IDF vectorizer

Initialize classifier $\boldsymbol{f}(\cdot) \in \{\boldsymbol{LR}, \boldsymbol{RF}, \boldsymbol{SVM}\}$

Preprocess sentence s

Convert to lowercase

Remove punctuation and non-alphabetic characters

Tokenize and lemmatize

Compute lexical feature vector $x_{tfidf}(s)$

Compute semantic embedding xsbert(s) ∈ R^{384}

Normalize feature vectors

$$\widehat{x_{tfidf}} = \frac{x_{tfidf}}{\|x_{tfidf}\|2} \quad (10)$$

$$\widehat{x_{SBERT}} = \frac{x_{SBERT}}{\|x_{SBERT}\|2} \quad (11)$$

Construct hybrid representation

$$xhybrid(s) = \hat{x}tfidf \oplus \hat{x}sbert$$

If training phase then

Apply SMOTE on hybrid training vectors

Train classifier f (x_{hybrid})

End

Compute prediction score

$$\boldsymbol{score = f(xhybrid)}$$

If score ≥ 0 then

y^=1

Else

y^=0

End

Return y^

F. Evaluation Metrics

Model performance was assessed using standard classification metrics. Accuracy is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

Precision and recall are computed as:

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN} \quad (13)$$

The F1-score, representing the harmonic mean of precision and recall, is calculated as:

$$F1 - score = \frac{2*Recall*precision}{Recall+precision} \quad (14)$$

Overall, the research design systematically evaluates lexical, semantic, and hybrid feature representations under controlled experimental conditions, ensuring that performance differences arise from feature learning strategies rather than inconsistencies in preprocessing, data splitting, or model configuration.

G. Computational Overhead and Deployment Considerations

From a practical deployment perspective, the hybrid framework involves two stages of feature extraction with distinct computational profiles. TF-IDF vectorization is lightweight, requiring only term-frequency counting over the vocabulary, and scales linearly with corpus size. SBERT inference using the all-MiniLM-L6-v2 model requires a transformer forward pass per sentence; on a standard CPU, average inference latency is approximately 5–10 milliseconds per sentence, making it feasible for batch processing in software development tool chains. The SVM classifier, operating on the concatenated high-dimensional feature vector, adds negligible inference overhead. The total pipeline can be integrated into requirements engineering tools as an offline batch analyzer or as a lightweight service for interactive requirement review, with memory requirements well within the constraints of modern development environments.

IV. RESULTS

This section presents the experimental evaluation of the proposed ambiguity detection framework on the Functional–Non-Functional Requirements (FR–NFR) dataset. The dataset contains 6008 software requirement statements, out of which 5665 are labelled as clear and 343 as ambiguous, indicating a severe class imbalance. All models were evaluated using accuracy, precision, recall, F1-score, and overall score. Particular emphasis is placed on recall and F1-score because ambiguity detection is inherently an imbalanced classification problem.

Three classical machine-learning classifiers were selected: Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM). Each classifier was evaluated using three feature configurations: TF-IDF, SBERT, and the hybrid TF-IDF + SBERT representation. The purpose of this evaluation is to systematically analyze how lexical and semantic representations influence ambiguity detection under consistent modelling conditions.

A. Baseline TF-IDF Results

Baseline experiments were conducted using TF-IDF lexical features. The results show that TF-IDF-based models achieve moderate to high accuracy; however, their effectiveness in detecting ambiguous requirements remains limited. Logistic Regression achieved an accuracy of 0.760355 with a recall of 0.584416, but precision was very low (0.133531), resulting in a weak F1-score of 0.217391. Similarly, SVM with TF-IDF obtained a recall of 0.629870, but precision dropped to 0.120497, producing an F1-score of only 0.202294. Random Forest achieved higher accuracy (0.925666), but recall remained moderate (0.454545), indicating that many ambiguous requirements were still missed.

These findings demonstrate that lexical frequency features alone are insufficient for reliable ambiguity detection. Although TF-IDF captures word importance, it does not model contextual relationships between words, leading to unstable precision–recall trade-offs under class imbalance.

B. SBERT Results

To evaluate the impact of contextual semantic representation, SBERT embeddings were used with LR, RF, and SVM classifiers. SBERT-based models showed improved contextual understanding compared to TF-IDF-only baselines. Logistic Regression achieved an F1-score of 0.455100 with a recall of 0.493500, while SVM achieved an F1-score of 0.445100 with a recall of 0.461000. These values indicate better balance compared to lexical-only models.

However, Random Forest trained on SBERT embeddings achieved very high accuracy (0.953800) and precision (0.837200), but recall dropped to 0.233800, producing an F1-score of 0.365500. This behavior indicates a strong bias toward the majority class, where ambiguous requirements are rarely predicted.

These results confirm that although SBERT captures contextual meaning and sentence-level semantics, semantic embeddings alone do not provide stable ambiguity detection under severe class imbalance.

C. Hybrid TF-IDF + SBERT Results

The proposed hybrid framework integrates lexical TF-IDF features with semantic SBERT embeddings into a unified representation. The objective of this design is to capture complementary information that cannot be effectively modelled using single-feature configurations. All hybrid models were evaluated using LR, RF, and SVM.

The experimental results demonstrate a clear and consistent improvement when hybrid features are applied. SVM with TF-IDF + SBERT achieved the best overall performance, obtaining an accuracy of 0.970400, precision of 0.798400, recall of 0.642900, and an F1-score of 0.712200, resulting in the highest overall score of 0.780975. Logistic Regression with hybrid features also demonstrated strong performance, achieving an F1-score of 0.684000. Random Forest benefited from hybrid features in terms of precision (0.833300), but recall remained low (0.259700), limiting its F1-score.

The improvement is primarily observed in recall and F1-score, indicating better detection of ambiguous requirements. These results confirm that ambiguity detection requires simultaneous modelling of lexical frequency patterns and contextual semantic information. The full comparison of all classifiers and feature configuration combinations is presented in Table I.

TABLE I. PERFORMANCE COMPARISON ON FR–NFR DATASET

Feature Set	Model	Accuracy	Precision	Recall	F1-Score	Overall Score
TF-IDF	SVM	0.717086	0.120497	0.629870	0.202294	0.417437
TF-IDF	LR	0.760355	0.133531	0.584416	0.217391	0.423923
TF-IDF	RF	0.925666	0.374332	0.454545	0.410557	0.541275
SBERT	RF	0.953800	0.837200	0.233800	0.365500	0.597575
SBERT	LR	0.932700	0.422200	0.493500	0.455100	0.575875
SBERT	SVM	0.934500	0.430300	0.461000	0.445100	0.567725
TF-IDF + SBERT	SVM	0.970400	0.798400	0.642900	0.712200	0.780975
TF-IDF + SBERT	LR	0.968600	0.800000	0.597400	0.684000	0.762500
TF-IDF + SBERT	RF	0.954900	0.833300	0.259700	0.396000	0.610975

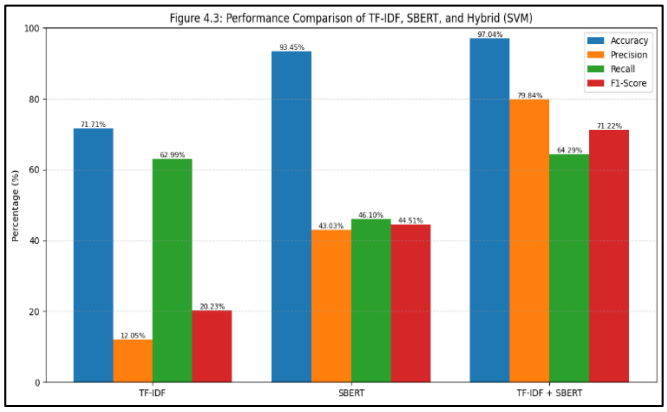


Fig. 2. Performance comparison of TF-IDF, SBERT, and a hybrid model.

V. Analysis and Discussion

The experimental results reveal significant differences between single-feature and hybrid feature representations. In the FR–NFR dataset, class imbalance strongly influences model behavior. Baseline TF-IDF models achieve moderate to high accuracy due to the dominance of clear requirements, yet they fail to consistently detect ambiguous requirements, as reflected in low precision and F1-scores.

SBERT embeddings improve contextual understanding and produce more balanced precision–recall behavior compared to TF-IDF. However, when used independently, SBERT remains sensitive to class imbalance, particularly in tree-based models such as Random Forest, which show conservative prediction behavior. Previous studies have similarly observed that contextual embeddings enhance requirement classification but require integration with complementary features for stable minority-class detection [15], [19].

The hybrid TF-IDF + SBERT representation consistently improves recall and F1-score across classifiers. This confirms that ambiguity is a multi-dimensional phenomenon requiring both lexical and semantic modelling. Lexical features capture frequency-based patterns associated with vague terms, while SBERT embeddings encode contextual meaning and sentence intent. Their integration reduces false negatives and produces more stable classification boundaries.

Support Vector Machine consistently achieves the best performance under a hybrid configuration. Its margin-based learning mechanism allows effective separation within the high-dimensional fused feature space. Unlike Logistic Regression, which tends to favor the majority class, and Random Forest, which remains conservative, SVM maintains a better balance between precision and recall.

Overall, the results demonstrate that hybrid feature integration provides a robust and generalizable solution for ambiguity detection in functional–non-functional requirements. The consistent improvement in recall and F1-score confirms the practical utility of the proposed framework, particularly when combined with Support Vector Machine. The cross-classifier performance trends are further visualized in Fig. 2, which illustrates the F1-score comparison across all feature configurations and classifiers. The obtained F1-score reflects the inherent difficulty of ambiguity detection under severe class imbalance, where minority-class instances are limited and context-dependent.

A key limitation of the current evaluation is that the framework is assessed on a single dataset from one source. Without cross-dataset validation or transfer experiments on requirements from different application domains, it remains uncertain whether the observed performance generalizes beyond the FR–NFR dataset. Future work should evaluate the framework on additional publicly available requirements corpora to establish broader generalizability. Furthermore, while the hybrid approach outperforms standalone TF-IDF and SBERT baselines, a direct comparison against fine-tuned transformer models such as BERT or RoBERTa on the same dataset would provide stronger evidence of the framework's competitive standing relative to the current state-of-the-art [23]. Due to computational constraints and the focus on feature-level hybrid representation, fine-tuned transformer baselines such as BERT and RoBERTa were not included in the current experimental evaluation.

VI. Conclusion and Future Work

This study investigated ambiguity detection in software requirements using a hybrid feature learning framework that integrates TF-IDF and SBERT representations within classical machine learning classifiers. Ambiguous requirements remain a major source of project delay and miscommunication, and traditional lexical models are often insufficient to capture context-dependent ambiguity.

The experimental results on the FR–NFR dataset show that single-feature models such as TF-IDF or SBERT, when used independently, produce unstable recall and limited F1-scores under severe class imbalance. TF-IDF captures lexical importance but lacks contextual understanding, while SBERT provides semantic awareness but struggles when applied alone. These findings confirm that ambiguity is multi-dimensional and cannot be reliably detected through isolated feature representations.

The hybrid TF-IDF + SBERT configuration consistently achieved the strongest and most balanced performance, particularly when evaluated using Support Vector Machine. Improvements were most evident in recall and F1-score, indicating a reduction in false negatives, which is critical in ambiguity detection tasks. The results demonstrate that integrating lexical and semantic cues within a unified representation significantly enhances classification robustness and minority-class detection.

Overall, the proposed hybrid framework provides a reproducible and empirically validated solution for ambiguity detection, outperforming single-feature baselines and offering practical applicability in requirements engineering.

A. Limitations

The study is limited to English-language requirements and focuses on sentence-level ambiguity without modelling document-level context. The framework is evaluated on a single dataset from one source; cross-dataset validation across requirements from different application domains has not been performed, which limits conclusions about generalizability.

Although the FR–NFR dataset is substantial, it may not fully represent the linguistic diversity of large industrial specifications. The proposed hybrid approach is not directly compared against fine-tuned transformer-based baselines such as BERT or RoBERTa on the same dataset, which would be necessary to position the framework conclusively relative to the current state-of-the-art. In addition, SBERT embeddings were not domain-adapted to software engineering corpora.

B. Future Work

Future work may extend this framework to larger and multilingual datasets to improve generalizability, and should include cross-dataset validation on requirements from diverse application domains. A direct comparison against fine-tuned transformer-based models such as BERT and RoBERTa on the FR–NFR dataset would clarify the competitive standing of the hybrid approach relative to the current state-of-the-art [23]. Domain-specific fine-tuning of transformer embeddings could further enhance contextual sensitivity. Incorporating explainable AI techniques such as SHAP or LIME would improve transparency and stakeholder trust. Finally, document-level modelling approaches could enable detection of cross-requirement ambiguity, strengthening real-world applicability.

REFERENCES

[1] A. Birhane, "The Impossibility of Automating Ambiguity," Artif. Life, vol. 27, no. 1, pp. 44–61, 2021, doi: 10.1162/artl_a_00336.

[2] M. Q. Riaz, W. H. Butt, and S. Rehman, "Automatic Detection of Ambiguous Software Requirements: An Insight," in 2019 5th International Conference on Information Management (ICIM), 2019, pp. 1–6. doi: 10.1109/INFOMAN.2019.8714682.

[3] M. Yaseen, A. Mustapha, M. A. Shah, and N. Ibrahim, "A hybrid technique using minimal spanning tree and analytic hierarchical process to prioritize functional requirements for parallel software development," Requir. Eng., vol. 28, no. 3, pp. 347–376, Feb. 2023, doi: 10.1007/s00766-023-00397-9.

[4] O. Kosenkov et al., "Systematic Mapping Study on Requirements Engineering for Regulatory Compliance of Software Systems," 2024. doi: 10.48550/arXiv.2411.01940.

[5] M. Yaseen and Z. Karamat, "Requirements Engineering Model (REM): An Assessment Model for Software Vendor Organizations," J. Softw. Evol. Process, vol. 37, no. 4, Apr. 2025, doi: 10.1002/smr.70020.

[6] A. Veizaga, S. Y. Shin, and L. C. Briand, "Automated Smell Detection and Recommendation in Natural Language Requirements," IEEE Trans. Softw. Eng., vol. 50, no. 4, pp. 695–720, 2024, doi: 10.1109/TSE.2024.3361033.

[7] S. K. Aqsa Rasheed, Bushra Zafar, Tehmina Shehryar, Naila Aiman Aslam, Muhammad Sajid, Nouman Ali, Saadat Hanif Dar, "Requirement Engineering Challenges in Agile Software Development," 2021.

[8] S. Ezzini, S. Abualhaija, C. Arora, and M. Sabetzadeh, "Automated handling of anaphoric ambiguity in requirements: a multi-solution study," in Proceedings of the 44th International Conference on Software Engineering, in ICSE '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 187–199. doi: 10.1145/3510003.3510157.

[9] S. Ezzini, S. Abualhaija, C. Arora, M. Sabetzadeh, and L. C. Briand, "Using Domain-Specific Corpora for Improved Handling of Ambiguity in Requirements," in 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), 2021, pp. 1485–1497. doi: 10.1109/ICSE43902.2021.00133.

[10] M. Yaseen, A. Bahar, G. Rahman, M. Ashurov, and E. Kholiyarov, "Minimizing Inter-Dependencies in Functional Requirements for Timely Delivery of Software Projects: A Prioritization Approach Using AHP and Spanning Trees," J. Softw. Evol. Process, vol. 37, no. 11, p. e70063, 2025, doi: https://doi.org/10.1002/smr.70063.

[11] Y. Liu, A. Medlar, and D. Głowacka, "Lexical ambiguity detection in professional discourse," Inf. Process. Manag., vol. 59, no. 5, p. 103000, 2022, doi: https://doi.org/10.1016/j.ipm.2022.103000.

[12] A. Ferrari et al., "Detecting requirements defects with NLP patterns: an industrial experience in the railway domain," Empir. Softw. Eng., vol. 23, no. 6, pp. 3684–3733, 2018, doi: 10.1007/s10664-018-9596-7.

[13] B. Gleich, O. Creighton, and L. Kof, "Ambiguity Detection: Towards a Tool Explaining Ambiguity Sources," in Requirements Engineering: Foundation for Software Quality, R. Wieringa and A. Persson, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 218–232.

[14] M. Alhamed and T. Storer, "Evaluation of Context-Aware Language Models and Experts for Effort Estimation of Software Maintenance Issues," in 2022 IEEE International Conference on Software Maintenance and Evolution (ICSME), 2022, pp. 129–138. doi: 10.1109/ICSME55016.2022.00020.

[15] A. Moharil and A. Sharma, "Identification of intra-domain ambiguity using transformer-based machine learning," in Proceedings of the 1st International Workshop on Natural Language-based Software Engineering, New York, NY, USA: ACM, May 2022, pp. 51–58. doi: 10.1145/3528588.3528651.

[16] F. Yucalar, "Developing an Advanced Software Requirements Classification Model Using BERT: An Empirical Evaluation Study on Newly Generated Turkish Data," Appl. Sci., vol. 13, no. 20, 2023, doi: 10.3390/app132011127.

[17] K. Kaur and P. Kaur, "Improving BERT model for requirements classification by bidirectional LSTM-CNN deep model," Comput. Electr. Eng., vol. 108, p. 108699, 2023, doi: https://doi.org/10.1016/j.compeleceng.2023.108699.

[18] S. A. H. Bahtiar, I. Universitas Islam Indonesia, C. K. Dewa, I. Universitas Islam Indonesia, A. Luthfi, and I. Universitas Islam Indonesia, "Comparison of Naïve Bayes and Logistic Regression in Sentiment Analysis on Marketplace Reviews Using Rating-Based Labeling," 2023.

[19] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," CoRR, vol. abs/1908.1, 2019.

[20] N. Rankovic and D. Rankovic, "Labeling issues through semantic pattems in open-source Agile practices," Knowledge-Based Syst., vol. 328, p. 114197, 2025, doi: https://doi.org/10.1016/j.knosys.2025.114197.

[21] L. Zhao et al., "Natural Language Processing for Requirements Engineering: A Systematic Mapping Study," ACM Comput. Surv., vol. 54, no. 3, Apr. 2021, doi: 10.1145/3444689.

[22] R. Sonbol, G. Rebdawi, and N. Ghneim, "The Use of NLP-Based Text Representation Techniques to Support Requirement Engineering Tasks: A Systematic Mapping Review," IEEE Access, vol. 10, pp. 62811–62830, 2022, doi: 10.1109/ACCESS.2022.3182372.

[23] A. Ferrari and A. Esuli, "An NLP approach for cross-domain ambiguity detection in requirements engineering," Autom. Softw. Eng., vol. 26, pp. 559–598, 2019, doi: 10.1007/s10515-019-00261-7.

[24] S. S. T. Sonali, "FR-NFR Dataset," 2024, [Online]. Available: https://data.mendeley.com/datasets/4ysx9fyzv4/1/files/15b2b7dd-15d3-4db1-a510-1011fb1465a1

[25] G. Giray, K. E. Bennin, Ö. Köksal, Ö. Babur, and B. Tekinerdogan, "On the use of deep learning in software defect prediction," J. Syst. Softw., vol. 195, p. 111537, 2023, doi: https://doi.org/10.1016/j.jss.2022.111537.

QR Code-Based Access Control Systems: Architectural Taxonomy, Security Landscape, and Future Research Directions

Worood Alsawi, Dina M. Ibrahim
Department of Information Technology-College of Computer, Qassim University, Buraydah, Saudi Arabia

***Abstract*—Access control systems secure physical and digital settings, especially in colleges, businesses, and restricted areas. Physical key, magnetic card, and biometric access control techniques have loss, duplication, high deployment costs, and maintenance complexity. With the widespread adoption of smartphones, QR code-based access control systems (QR-ACS) have emerged as a flexible and cost-effective alternative. Scannable QR codes allow fast authentication without hardware, boosting user ease. How QR codes are generated, handled, and checked in the system determines the success of QR-based access control. This systematic review examines the integration and evolution of QR Code-Based Access Control Systems (QR-ACS), with particular attention to both recent innovations and the challenges that continue to accompany their adoption. A broad set of studies published between 2015 and 2023 was reviewed in order to explore how these systems have been designed, implemented, and evaluated across different application contexts. The analysis draws on literature indexed in major academic databases, including IEEE Xplore, ACM Digital Library, and JSTOR, with an emphasis on system architecture, implementation strategies, and reported performance outcomes. Overall, the reviewed studies indicate that QR-ACS can enhance operational efficiency and offer practical security benefits, especially when combined with complementary technologies. At the same time, recurring concerns related to security, robustness, and deployment limitations remain evident. This systematic review analyzes QR-ACS architectures, security mechanisms, and threat models, with particular emphasis on IoT integration, authentication strategies, and risk-aware system design.**

Keywords*—*QR codes; access control systems; security mechanisms; authentication; Internet of Things

I. Introduction

In the realm of security technology, the adoption of QR Code-Based Access Control Systems (QR-ACS) has been transformative, offering a novel method to manage access to various physical and digital spaces. QR codes, originally designed for tracking parts in vehicle manufacturing, have evolved significantly, becoming pivotal in access control systems due to their simplicity, cost-effectiveness, and ease of use. As Adedoyin and Olukoya [1] note, "QR codes, once developed for industrial purposes, now support smart lock systems and various security solutions," demonstrating their adaptability across multiple domains. These systems utilize QR codes to authorize entry to secured areas, integrating seamlessly with mobile technology and cloud computing to provide dynamic and robust security solutions [2]. This evolution underscores the ability of QR codes to transcend their original purpose, becoming a versatile tool for a wide range of applications in modern security systems.

As global security challenges continue to evolve, the demand for access control solutions that can adapt to changing operational conditions has become more apparent. Conventional approaches, including keycards, passwords, and biometric systems, are widely used but often present practical limitations, such as vulnerability to loss or theft and the need for dedicated infrastructure. In this context, QR-ACS are frequently discussed as a more flexible and cost-efficient option. According to Petrea et al. (2020), these systems can reduce operational overhead by relying on mobile devices and cloud-based services, which makes them suitable for environments where physical credentials or complex hardware are difficult to manage [3]. The widespread availability of smartphones with built-in QR code scanning capabilities has further supported the adoption of QR-ACS across both public and private sectors. Beyond conventional access scenarios, these systems have also been applied in situations that require minimal physical interaction. Awotunde et al. [4] highlights the growing relevance of such contactless solutions, particularly during health-related crises such as the COVID-19 pandemic, where reducing physical contact became a key safety consideration.

Several sectors, including education, healthcare, transportation, and corporate offices, have begun to integrate QR-ACS into their security frameworks. In educational settings, for instance, QR codes are increasingly used to manage access to facilities such as libraries, laboratories, and examination halls. Similarly, healthcare environments have adopted QR codes to control access to patient records, restricted areas, and medical equipment, particularly in cases where traditional physical access mechanisms may be inefficient or introduce additional risks. At the same time, security requirements differ across these contexts, as certain areas—such as research laboratories, server rooms, and administrative offices—demand higher levels of protection. While the adaptable nature of QR-ACS allows them to support varying access levels, further research remains necessary to refine these systems and ensure that they can meet more demanding security needs. Moreover, as organizations increasingly shift toward IoT-enabled environments, the combination of QR-ACS with biometric data and real-time monitoring has received growing attention as a potential approach to strengthening access control frameworks [6].

For example, [7] proposed an access control system for smart buildings that employs aesthetic QR codes combined with symmetric encryption algorithms. In this system, QR codes incorporate user photographs as an additional authentication layer. Although the proposed solution was described as cost-effective and flexible, it exhibited limitations related to scalability and reliability, particularly in low-light conditions, which negatively affected the QR code scanning process.

In a related work, [8] presented a QR code-based smart door access control system designed for educational and corporate environments. The system enhanced operational efficiency by recording user access times through a centralized database; however, the authors reported that the solution still faced constraints related to system scalability and long-term deployment in larger environments.

This review seeks to synthesize and critically examine the existing literature on QR-ACS, with the aim of identifying both the technological advances that have addressed key challenges and the areas where notable limitations remain. Through the analysis of a diverse body of studies, the review explores the reported strengths and weaknesses of current QR-ACS implementations, as well as the approaches that appear most promising for future development. Particular attention is given to application contexts in which QR-ACS have demonstrated practical effectiveness, alongside scenarios where further refinement is still required. By addressing these aspects, the review contributes to a clearer understanding of the current role of QR-ACS in modern security environments and highlights potential research directions that may support improvements in security, efficiency, and scalability [9]. The main contributions of this paper are:

- A comprehensive systematic review of QR-ACS (2015–2024).
- A structured architectural taxonomy of QR-ACS.
- A comparative analysis across multiple application domains.
- A proposed risk assessment framework for QR-based security threats.
- Identification of research gaps and future directions.

II. Background

Access control systems play a central role in security management, as they define who is permitted to access a given resource, whether a physical environment, such as a building or a digital asset, such as sensitive data. Early access control mechanisms were based on simple physical solutions, including locks and keys. Over time, these approaches evolved into more complex electronic systems that incorporate cards, biometric identifiers, and, more recently, mobile-based technologies. As noted by Norman (2011), the transition from mechanical to electronic access control represented an important step toward greater flexibility and scalability in security management [10].

Fig. 1 and Fig. 2 illustrate the technological evolution and comparative characteristics of QR codes within the broader landscape of two-dimensional barcode technologies. Fig. 1 presents a conceptual comparison between QR codes and other matrix codes, such as PDF417, Data Matrix, and MaxiCode, highlighting their advantages in terms of data capacity, printing efficiency, and high-speed readability. Meanwhile, Fig. 2 outlines the chronological development of QR code technology, beginning with its invention in 1994 and progressing through key phases of adoption, including its expansion into marketing and payment systems, integration with IoT environments, and its recent role in advanced security applications such as biometric authentication, blockchain-based systems, and dynamic QR technologies. Together, these figures demonstrate how QR codes have evolved from industrial tracking tools into a foundational component of modern digital security and access control infrastructures.

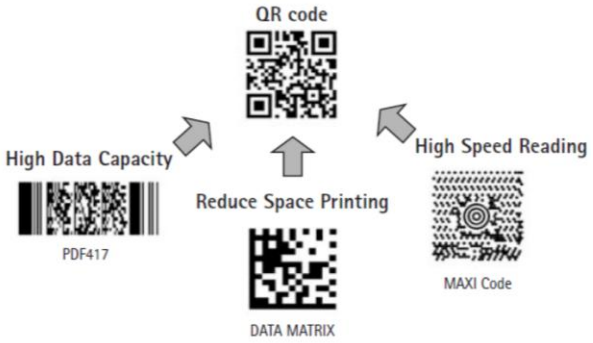


Fig. 1. Evolution of QR codes from manufacturing tracking to modern access control applications [5].



Fig. 2. Evolution of QR code technology.

Despite their advantages, QR-ACS are associated with a number of challenges, including the risk of QR code replication, unauthorized sharing, and potential security breaches resulting from code interception. In their basic form, QR codes do not provide built-in encryption or authentication mechanisms, which makes them vulnerable to different types of attacks. In response to these limitations, several enhanced QR-ACS solutions have been proposed that integrate additional security measures, such as encryption techniques, time-limited QR codes, and multi-factor authentication. These approaches aim to reduce security risks by limiting unauthorized access while preserving the flexibility and usability that characterize QR code-based systems. Nevertheless, the reported effectiveness of such solutions is not consistent across studies, and their performance often depends on the application context and implementation design. As a result, further research remains

necessary to establish standardized and scalable security frameworks that can be reliably applied across different sectors.

The growing research interest in QR-ACS is illustrated in Fig. 3. The figure presents the publication trend of QR-ACS-related research between 2015 and 2024. The data demonstrate a steady growth trajectory, with a noticeable acceleration after 2019, coinciding with the expansion of IoT ecosystems and increased demand for contactless authentication solutions during the COVID-19 pandemic. The upward trend in 2023–2024 suggests sustained research interest, particularly in areas integrating encryption, blockchain, and biometric authentication. This growth pattern confirms that QR-ACS remains an active and evolving research domain, warranting systematic synthesis and structured analysis.

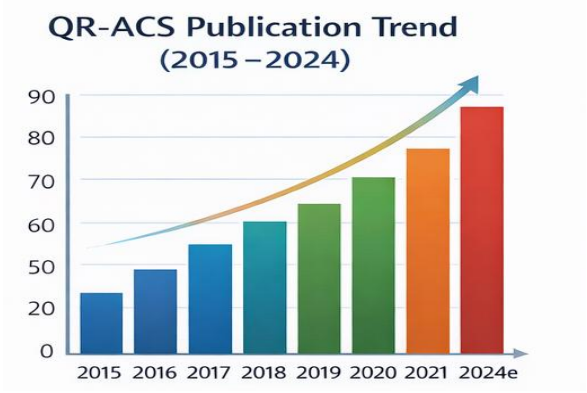


Fig. 3. QR-ACS publication trend (2015–2024).

At a fundamental level, the objective of an access control system is to enforce predefined access policies that protect facilities and information from unauthorized use. The emergence of electronic access control systems during the 1960s marked a significant shift away from purely mechanical solutions. Early implementations relied on magnetic stripe cards to support more dynamic authorization processes, which were later extended through the adoption of RFID (Radio Frequency Identification) technology. According to Rieback, Crispo, and Tanenbaum (2006), RFID-based systems introduced contactless operation alongside encrypted authentication data, thereby reshaping conventional access control mechanisms [11].

Subsequent developments in RFID-based access control further enhanced security capabilities by enabling the integration of authentication, tracking, and monitoring functions within a unified framework [12]. These developments illustrate how access control technologies have continued to adapt in response to increasing security requirements and the growing complexity of modern infrastructures. An overview of access control system architectures and their relationship to QR code-based solutions is illustrated in Fig. 4.

QR codes, commonly referred to as Quick Response codes, were introduced in 1994 by the Japanese company Denso Wave for the purpose of tracking components in vehicle manufacturing. As noted by de Seta (2023), Denso Wave initially developed QR code technology to improve efficiency in automotive production, which later supported its broader adoption [14]. Unlike conventional one-dimensional barcodes, QR codes are two-dimensional and were designed to encode information more efficiently while allowing faster scanning. This design enables QR codes to store various types of data, including URLs, contact details, and encrypted information, contributing to their versatility across different applications [15].



Fig. 4. Overview of access control systems, highlighting QR code-based solutions [13].

The widespread use of QR codes expanded further with the growth of smartphone technology. Smartphones equipped with built-in cameras have played a key role in facilitating quick and convenient QR code scanning in everyday contexts. As explained by Pandey and Awasthi (2020), this capability has supported the adoption of QR codes across multiple domains, ranging from marketing and digital payments to information sharing [16]. In the context of access control systems, QR codes are commonly used to store access credentials or to reference secure credentials maintained in cloud-based platforms. This approach has been described as a practical and scalable solution for managing access in diverse environments, including corporate offices, public venues, and residential settings [17].

Fig. 5 illustrates the structural components of a typical QR code and the arrangement of its functional elements. These components include position markers that assist scanners in detecting the orientation of the code, data modules that store encoded information, and quiet zones that ensure reliable recognition during scanning. Understanding the internal structure of QR codes is essential for explaining how they support efficient data encoding, error correction, and rapid decoding in access control applications.

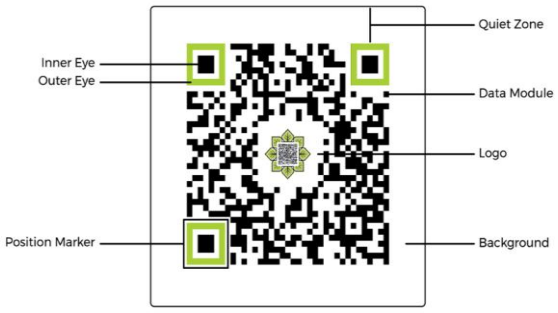


Fig. 5. Structure of a QR code, highlighting key components such as data modules, timing patterns, and error correction blocks [18].

III. Research Methodology

This study follows a systematic review methodology based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework to ensure transparency, reproducibility, and rigor in the literature selection process.

A. Data Sources and Search Strategy

A comprehensive literature search was conducted across multiple academic databases, including IEEE Xplore, ACM Digital Library, Scopus, SpringerLink, and Google Scholar. The search focused on studies published between 2015 and 2024 to capture recent developments in QR Code-Based Access Control Systems (QR-ACS). The search queries were constructed using combinations of relevant keywords, including: "QR code access control", "QR authentication", "QR-based security systems", "QR code IoT access control", "secure QR authentication", and "QR-based identity verification". Boolean operators (AND, OR) were used to refine the search and ensure comprehensive coverage of relevant studies.

B. Inclusion and Exclusion Criteria

To ensure relevance and quality, predefined inclusion and exclusion criteria were applied during the screening process. Inclusion criteria:

- Studies published between 2015 and 2024.
- Peer-reviewed journal articles and conference papers.
- Studies focusing on QR code-based authentication or access control systems.
- Studies presenting technical, experimental, or system-level contributions.

Exclusion criteria:

- Non-English publications.
- Studies not directly related to access control or authentication.
- Review papers without technical contribution.
- Short papers, editorials, posters, or non-peer-reviewed content.

C. Screening and Selection Process

The study selection process followed the PRISMA 2020 framework to ensure transparency and reproducibility. As illustrated in Fig. 6, a total of 1,275 records were initially identified from multiple academic databases, including IEEE Xplore, ACM Digital Library, Scopus, SpringerLink, and Google Scholar. After removing duplicate records (n = 352), 923 studies remained for title and abstract screening. During this phase, 668 records were excluded due to irrelevance to QR code-based access control systems. The remaining 255 articles were subjected to full-text review, of which 28 could not be retrieved. A total of 227 studies were assessed for eligibility, leading to the exclusion of 161 papers based on predefined criteria, including lack of relevance, non-QR-based approaches, and non-empirical contributions. Ultimately, 66 studies were included in the final qualitative and quantitative synthesis.

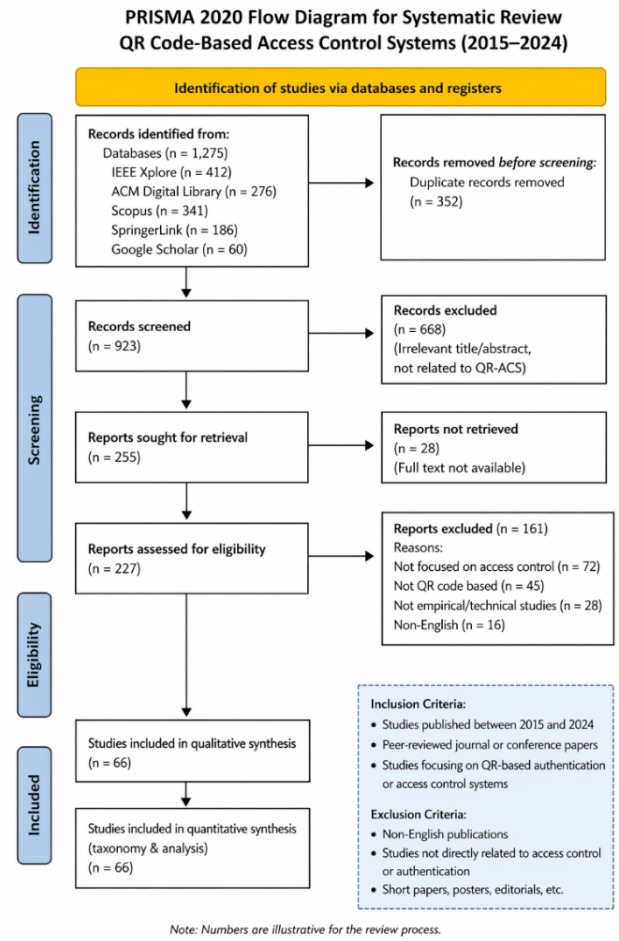


Fig. 6. PRISMA 2020 flow diagram illustrating the study selection process for the systematic review of QR Code-Based Access Control Systems (QR-ACS).

D. Data Extraction and Analysis

For each selected study, relevant data were systematically extracted, including system architecture, authentication mechanisms, security techniques, application domain, and reported limitations. The extracted studies were categorized into thematic groups to support comparative analysis and the development of a structured taxonomy. This categorization enabled the identification of architectural patterns, security trends, and research gaps within QR-ACS.

Additionally, a comparative evaluation was conducted to analyze trade-offs between security strength, system complexity, and deployment feasibility across different QR-ACS implementations.

IV. Literature Review

This literature review examines a broad range of studies that apply QR code-based access control systems across different application domains, including security and surveillance, mobile and IoT systems, transportation and ticketing, and healthcare. The reviewed work reflects the growing interest in QR code technology as a means of supporting authentication, enhancing security, and improving operational efficiency in varied contexts.

Within the area of security and surveillance, several studies investigate the integration of QR codes with complementary technologies such as facial biometrics, encryption techniques, and real-time monitoring. These approaches are generally proposed as alternatives to traditional access control mechanisms, including physical keys and swipe cards, with the intention of improving convenience while maintaining acceptable security levels.

Research focusing on mobile and IoT systems often emphasizes the practical advantages of QR code-based frameworks, particularly their cost efficiency, flexibility, and compatibility with smartphones and IoT infrastructures. These characteristics are frequently cited as factors that enable more streamlined access control and improved user interaction. In a similar manner, studies related to transportation and ticketing explore the use of encrypted QR codes to strengthen e-ticket security, while also acknowledging challenges associated with system interoperability and vulnerability management.

In the healthcare domain, the literature highlights the use of QR codes for controlling access to sensitive patient data and restricted areas, with particular attention given to encryption methods and authentication mechanisms required to protect electronic health records.

Across these application areas, QR code technology is commonly described as adaptable and capable of supporting a range of access control requirements. At the same time, recurring challenges are noted in the literature, including concerns related to scalability, the potential for unauthorized QR code duplication, and the need for supporting technical infrastructure. Taken together, these observations provide a basis for understanding the current use of QR code-based access control systems and point to several areas where further research is needed, especially with respect to security robustness and system interoperability.

A. Security and Surveillance

This section examines the implementation of QR code-based systems in a range of security and surveillance applications. The studies discussed here consider how QR codes are combined with complementary technologies, such as biometric authentication, encryption algorithms, and real-time monitoring, to support access control objectives. Rather than addressing a single use case, the reviewed work covers different application environments and illustrates how QR code-based solutions have been employed to modernize access control mechanisms across multiple contexts.

To enhance system security, QR code-based access control is often integrated with additional security mechanisms such as encryption techniques, biometric authentication, centralized servers, and monitoring components. These technologies work together to ensure secure identity verification, protect transmitted data, and enable continuous monitoring of system activities.

Fig. 7 illustrates a conceptual architecture of QR code-based security and surveillance systems, highlighting the integration of QR authentication with encryption methods, biometric verification, logging mechanisms, and IoT-based monitoring components.

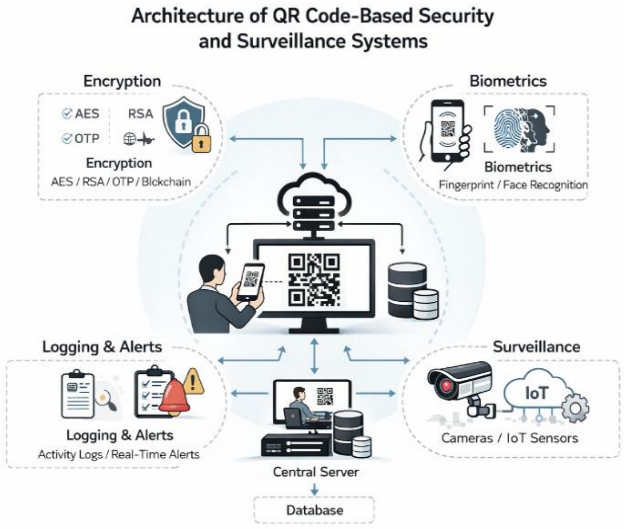


Fig. 7. Architecture of QR code–based security and surveillance systems illustrating the integration of QR authentication with encryption, biometric verification, centralized servers, and monitoring mechanisms.

1) Home security: Home security systems that incorporate QR code technology are often presented as alternatives to traditional key-based solutions, with an emphasis on improved convenience and flexible access management. In many proposed implementations, QR codes are integrated with additional technologies, such as facial recognition and IoT-based automation, to support real-time monitoring and remote control capabilities. These combinations are intended to strengthen residential security while reducing reliance on physical keys, although their effectiveness may vary depending on system design and deployment conditions. Table I summarizes representative studies that apply qr code-based and related technologies to home security scenarios, highlighting common design choices, reported benefits, and frequently noted limitations.

2) Office and workplace security: In office and workplace environments, QR code-based access control systems are increasingly considered as alternatives to traditional authentication methods, such as RFID cards, particularly when combined with encrypted QR codes and multi-factor authentication mechanisms. These systems are commonly proposed to support controlled access to sensitive areas, while also facilitating functions such as monitoring and attendance tracking. Table II summarizes representative studies that examine QR code-based access control solutions in office and workplace settings, highlighting proposed authentication approaches alongside their reported benefits and limitations.

TABLE I. STUDIES RELATED TO HOME SECURITY

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[19]	2014	Residential multi-factor security system	Residential multi-factor security system	Unauthorized entry	Cost-effective multi-layer security	NFC reliability; false PIR triggers
[20]	2022	Wireless home security system	RFID + Fingerprint + Face ID + ESP32-CAM	Weak traditional locks	Multi-factor secure access	System complexity; maintenance
[7]	2018	Smart building access with aesthetic QR	QR + Symmetric encryption + Photo verification	RFID cost; offline verification	Flexible, cost-effective access	Photo verification accuracy; scalability
[21]	2021	IoT-based smart home security	Face recognition (PCA) + IoT	Automated intrusion detection	Low-power real-time monitoring	Limited scalability testing
[22]	2020	Face recognition access control	OpenCV + LBPH + Raspberry Pi	Real-time authentication	Remote access with alerts	Limited scalability
[23]	2020	Biometric door access control	Fingerprint + Arduino + Bluetooth	Device interoperability	Reliable biometric access	Limited large-scale testing
[24]	2020	Dual biometric home security	Face (PCA) + Fingerprint	Unauthorized residential access	Strong biometric authentication	Lighting sensitivity; scalability
[25]	2022	IoT facial recognition security	Haar Cascade + LBPH + Mobile App	Home monitoring automation	92.86% recognition accuracy	Internet dependency; scalability

TABLE II. STUDIES RELATED TO OFFICE AND WORKPLACE SECURITY

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[26]	2022	Secure QR transmission for confidential data (e-tickets, coupons)	QR module modification + Error correction	Weak data protection in standard QR	Increased hidden data capacity; secure QR	Requires secret key access
[27]	2023	Device registration authentication system	QR + Fingerprint + C#/SQL	Paper receipt fraud; inefficient verification	Improved authentication and fraud reduction	User familiarity; hardware maintenance
[28]	2021	Residential IoT access control	QR + SHA-2 + IoT	Replacing physical keys	Improved security; real-time notifications	Scalability; internet dependency

3) University and research lab access control: In universities and research laboratories, QR codes are increasingly deployed in combination with technologies such as biometric authentication and encryption to support access control and attendance management. These approaches are commonly discussed as more flexible and cost-effective alternatives to conventional access control systems, particularly in environments that require frequent user verification and controlled access to specialized facilities. Table III summarizes representative studies that examine QR code-based access control solutions in university and research laboratory settings, highlighting a variety of cryptographic, biometric, and system-level approaches alongside their reported benefits and limitations.

TABLE III. STUDIES RELATED TO UNIVERSITY AND RESEARCH LAB ACCESS CONTROL

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[29]	2018	Secure access control using encrypted QR codes.	QR + RSA + PRNG	Unencrypted QR vulnerabilities.	Improved security; cross-platform compatibility	RSA overhead; higher device processing
[30]	2020	QR-based door access for university labs	QR + Raspberry Pi + Web server	Weak locks; RFID cost/complexity	Cost-effective access control	Preliminary testing; limited scalability
[31]	2017	Smart door security for restricted areas	Bluetooth + Android + Microcontroller	RFID and fingerprint weaknesses	Real-time monitoring; cost-effective	Limited range (9 m); user management issues
[32]	2023	Lab access control with QR and OTP	QR + RSA + AES + TOTP	Unauthorized access prevention	TOTP more secure than HOTP	Limited to lab environments
[8]	2023	QR-based attendance and access tracking	QR + Centralized DB	Manual errors; RF card dependency	Improved accuracy and cost efficiency	QR duplication risk; internet dependency
[33]	2020	QR-based access control with monitoring	QR + Lamport Hash Chain + Web	Security and monitoring challenges	Enhanced security and monitoring	Limited scalability; no usability feedback
[34]	2018	Biometric QR access for universities	QR + Face recognition + OpenCV	Authentication and privacy issues	Improved authentication and privacy	Limited scope; OpenCV constraints
[35]	2021	Secure login for e-exam systems	QR + SHA-256 + MD5	Brute force and unauthorized access	Secure and efficient login	Limited generalizability; missing usability data

4) Public spaces and government facilities: In government buildings and public spaces, QR code-based systems have been applied as part of access control strategies for managing high-traffic environments. These systems are commonly discussed as tools for supporting controlled access to sensitive areas, to reduce the risk of unauthorized entry while maintaining

efficient movement within public facilities. Table IV presents selected studies that address QR code-based access control in public spaces and government facilities, emphasizing emergency scenarios, decentralized authentication, and system-level considerations alongside reported limitations.

TABLE IV. STUDIES RELATED TO PUBLIC SPACES AND GOVERNMENT FACILITIES

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[36]	2018	Emergency access control	QR + Shamir Secret Sharing + ABE	No online authentication in emergencies	Secure offline attribute-based access	Attribute policy dependency; limited disaster testing
[37]	2020	Smartphone QR access control	QR + DES + Embedded system	Security and cost limits of locks/RFID/biometrics	Integrated access and attendance control	Smartphone dependency
[38]	2020	Unified identity platform	QR hybrid encryption + WSN + ZigBee	Fragmented identity systems	Low-power secure authentication	No real-world testing; scalability
[39]	2021	QR access control with monitoring	QR + Web app + ID scanner	QR duplication; setup complexity	Cost-effective access monitoring	Security risks; user adoption issues
[40]	2019	Blockchain QR authentication	QR + One-time QR + Blockchain	Secure decentralized authentication	Improved security via decentralized verification	Scalability; integration challenges

TABLE V. STUDIES RELATED TO HIGH-SECURITY ENVIRONMENTS

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[41]	2019	Biometric passport security	QR + AES + SHA-256 + Biometrics	RFID privacy risks	High-accuracy biometric authentication	Biometric reliability; infrastructure dependency
[42]	2015	LED-QR access authentication	QR + LED tags	Improving QR access security	Enhanced authentication performance	Scalability not evaluated
[43]	2020	Encrypted QR access control	QR + RSA + PRNG	Protecting authorization data	Improved data security	Key management; scalability
[44]	2024	Secure QR encryption for access control	QR + RSA + PRNG	Authorization data protection	Stronger encrypted access control	Limited real-world evaluation

5) High-security environments (e.g., defense or sensitive areas): In high-security environments, such as defense-related or sensitive government facilities, QR code-based systems are often discussed in conjunction with advanced encryption techniques and biometric integration to support stricter access control requirements. These approaches are proposed to manage access to highly restricted areas, where higher assurance levels are typically required compared to conventional access control settings. Table V summarizes studies that address QR code-based access control in high-security environments, including national security and sensitive infrastructure, with a focus on encryption and biometric integration.

In security and surveillance applications, QR code-based systems are commonly discussed as access control solutions that combine QR technology with elements such as encryption, real-time monitoring, and biometric verification. These systems are often proposed as flexible and user-friendly alternatives to traditional security mechanisms, including physical keys and RFID cards, particularly in contexts where cost and ease of deployment are important considerations. At the same time, the literature points to several challenges, including the susceptibility of QR codes to unauthorized duplication and the technical complexity associated with integrating additional security layers such as encryption. Accordingly, further research remains necessary to address these limitations and to assess the suitability of QR code-based systems for wider adoption in high-security environments

B. Mobile and IoT Systems

This section examines the application of QR code technology within mobile and IoT-based systems, with particular attention to its use in access control contexts. The reviewed studies discuss QR codes as cost-effective and flexible solutions that can be integrated with IoT devices to support secure access management. In many cases, these approaches are presented as a means of streamlining access processes while improving overall operational efficiency.

1) Smart home automation: In smart home environments, QR code-based systems are frequently combined with IoT devices to support automated security and access control functions. Such systems are commonly designed to allow homeowners to manage door locks, monitor entry points, and receive notifications through smartphone applications. These features are often highlighted as contributing to improved convenience and remote control capabilities, although their effectiveness may depend on system configuration and network reliability. Table VI summarizes studies that examine smart home automation systems integrating QR codes, IoT devices, and complementary technologies, highlighting proposed approaches, reported outcomes, and commonly noted limitations.

2) Industrial IoT applications: In industrial environments, QR code-based access control systems integrated with IoT technologies are commonly discussed as approaches for supporting security and operational efficiency. Such solutions

are applied to manage access to restricted areas and to assist in the monitoring of assets and personnel within industrial settings. These systems are often presented as part of broader industrial IoT frameworks, where access control is combined with real-time data collection and system management. Table VII highlights a representative approach to QR code-based authentication in industrial IoT contexts, emphasizing performance considerations alongside scalability challenges.

3) Mobile payment systems: In mobile payment systems, QR code-based authentication is commonly discussed as a practical approach for supporting secure and convenient financial transactions. Rather than relying solely on traditional password-based mechanisms, these systems often incorporate multi-factor authentication schemes, such as QR code verification combined with one-time passwords (OTPs). This combination is intended to enhance transaction security while maintaining ease of use for end users. Table VIII presents selected studies that examine the use of QR code-based authentication in mobile payment systems, highlighting multi-factor security approaches alongside practical limitations related to deployment and usability.

TABLE VI. STUDIES RELATED TO SMART HOME AUTOMATION

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[45]	2023	IoT smart door lock security	Bluetooth MAC + Raspberry Pi + Firebase	Weak traditional locks	Remote access and monitoring	Internet and smartphone dependency
[46]	2022	IoT facial recognition gate security	Face recognition + OpenCV + Raspberry Pi	Rising home break-ins	Real-time monitoring and alerts	Lighting and recognition accuracy
[47]	2022	IoT smart home gate access	Face recognition + Arduino + ESP8266	Centralized home security control	Remote monitoring via mobile	Network reliability; integration
[48]	2019	GPS-based smart door lock	GPS + STM32 + XBee	Manual lock limitations	Location-based automated access	GPS accuracy; connectivity
[49]	2020	IoT home automation security	Face detection + Raspberry Pi + IoT sensors	Automated home security	Real-time monitoring and alerts	Internet dependency; lighting issues
[50]	2021	QR access framework for Android	QR + ZXing + Encrypted ID	Lack of QR access frameworks	Flexible QR-based access system	Android focus; limited testing

TABLE VII. STUDIES RELATED TO INDUSTRIAL IOT APPLICATIONS

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[51]	2020	IoT QR authentication security	QRAM (3-layer QR authentication)	Fast and secure IoT authentication	Reduced computation time; improved security	High decryption complexity; scalability

TABLE VIII. STUDIES RELATED TO MOBILE PAYMENT SYSTEMS

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[52]	2010	QR-based user authentication	QR + OTP	Password-based security risks	Dynamic OTP authentication	Requires camera-enabled mobile device
[53]	2021	Secure mobile money authentication	QR + PIN + OTP + Fingerprint + SHA-256 + FIDO	Weak mobile money security	Improved protection against replay and phishing	Biometric device and infrastructure requirements
[54]	2023	Secure QR transactions for MSMEs	QR + GOST encryption	Secure QR payment transactions	Enhanced QR transaction security	Requires specialized QR reader
[55]	2010	Secure online banking authentication	QR + Mobile OTP + SSL/TLS	Phishing and weak authentication	Secure QR banking authentication	Limited testing; small sample

4) Smart Building and Facility Management: Smart buildings increasingly incorporate QR code-based systems as part of their access control, monitoring, and automation strategies. These solutions are commonly discussed in the literature as scalable approaches that support controlled entry to different areas within a building and facilitate integration with existing management systems. By enabling flexible access policies and centralized monitoring, QR code-based approaches are presented as contributing to more efficient facility management and improved security practices. Table IX highlights a representative application of QR code-based access control in smart building facilities, emphasizing shared access management and system scalability considerations.

TABLE IX. STUDIES RELATED TO SMART BUILDING AND FACILITY MANAGEMENT

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[56]	2019	Smart campus locker access	QR + IoT + Smart lock + Mobile app	Traditional locker limitations	Flexible shared locker access	Prototype stage; scalability

QR code-based access control in mobile and IoT systems is often discussed as an adaptable and scalable approach for securing connected devices and services. By relying on mobile platforms and IoT networks, these systems can support flexible access management across a range of environments, from residential automation to industrial settings. At the same time, the literature points to several ongoing challenges, including the need for reliable connectivity, compatibility across heterogeneous devices, and the diverse security requirements inherent to IoT ecosystems. As a result, future research is frequently directed toward the development of standardized frameworks that can facilitate cross-platform integration and support more robust and scalable QR code-based access control solutions within IoT environments.

C. Transportation and Ticketing Systems

In transportation and ticketing applications, QR code-based systems have been explored as a means of streamlining ticket verification processes and supporting secure transactions. Existing studies emphasize the use of encrypted QR codes to enable contactless ticketing, which can improve efficiency by automating procedures such as check-in and seat verification. These approaches are often associated with reductions in manual errors and improvements in operational workflows within public transportation systems. However, the literature also highlights practical challenges, including infrastructure requirements for managing biometric data and potential security concerns related to user-generated passwords in electronic ticketing systems. Addressing these issues may require further advancements in QR code encryption techniques, as well as improved system interoperability and deployment strategies. Table X summarizes representative studies that investigate the use of QR code-based authentication in transportation and ticketing systems, highlighting biometric integration, encrypted e-ticketing approaches, and associated deployment challenges.

TABLE X. STUDIES RELATED TO TRANSPORTATION AND TICKETING

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[57]	2019	Train ticketing and seat verification	QR + Fingerprint biometrics	Manual ticket verification errors	Automated check-in; improved security	Infrastructure requirements
[58]	2010	Secure QR e-ticket system	QR + Block cipher + Password	Password-based security risks	Improved ticket accessibility	Password theft; device compatibility

D. Healthcare Systems

In healthcare applications, QR code technology is primarily explored as a mechanism for securing access to sensitive patient information and supporting data privacy within electronic health record (EHR) systems. Existing studies investigate the use of encryption techniques and cloud-based storage to restrict data access to authorized personnel. At the same time, the literature identifies several challenges, including key management in large-scale deployments and the need for extensive testing to assess system performance in real-world healthcare settings. Overall, these studies suggest that QR code-based approaches have the potential to strengthen privacy protection in healthcare environments, while also indicating areas that require further development and validation. Table XI presents studies that examine the application of QR codes in healthcare systems, with a particular focus on securing access to sensitive patient information. The reviewed studies discuss encryption techniques and authentication frameworks that are intended to support data protection in electronic health record (EHR) systems.

TABLE XI. STUDIES RELATED TO HEALTHCARE SYSTEMS

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[59]	2022	Secure QR data transfer	QR + RSA	Personal data privacy	Secure encrypted data transfer	Key management complexity
[60]	2020	Secure EHR access via QR	QR + Cloud storage	Protecting sensitive health data	Secure EHR access framework	Limited real-world evaluation

E. Other Fields

In addition to the more established application domains, QR code-based systems have also been explored in a range of other environments, including hostel management, smart locker systems, and identity tracking applications. Studies in these areas describe how QR codes can be applied to support identity management, enhance security, and automate certain administrative processes. Alongside the reported advantages, such as reduced dependence on manual record-keeping, the literature also identifies several limitations. These include the need for reliable network connectivity and the possibility of scanning-related issues, particularly in settings with limited technical infrastructure. Overall, these studies reflect a growing research interest in extending the use of QR codes to diverse application scenarios beyond traditional access control systems. Table XII presents studies that examine the application of QR code technology in other fields, such as hostel management and smart building environments. The reviewed studies discuss how QR codes can be used to support identity management, improve security in residential and commercial contexts, and enable integration with cloud platforms and sensor networks for real-time monitoring.

The comparative analysis of QR Code-Based Access Control Systems (QR-ACS) reflects a wide range of applications that are shaped by specific security requirements and environmental conditions. The reviewed studies span multiple contexts, including educational institutions, emergency management scenarios, and data-sensitive environments, illustrating the flexibility of QR-ACS in addressing different access control needs.

TABLE XII. STUDIES RELATED TO OTHER FIELDS

Ref	Year	Focus Area	Algorithm	Challenges addressed	Findings	Limitations
[61]	2021	Hostel identity and access monitoring	Hostel identity and access monitoring	Manual monitoring inefficiency	Accurate movement tracking	Infrastructure dependency
[62]	2022	QR smart locker system	QR + ESP8266 + Smart lock	Traditional key limitations	Secure locker access	Response delay

In academic and institutional settings, QR-ACS are commonly applied to support identity verification and access management, often with the goal of reducing reliance on physical tokens and improving user convenience. By contrast, systems designed for emergency or critical situations tend to emphasize encryption-enhanced QR codes, allowing secure access to be maintained even when conventional infrastructure is unavailable.

From a technological perspective, many of the proposed systems integrate encryption techniques, such as RSA, and in some cases combine QR codes with biometric authentication to strengthen protection against unauthorized access. While these approaches contribute to improved security, the literature also consistently reports challenges related to QR code tampering, unauthorized duplication, system scalability, and the complexity of implementing advanced cryptographic mechanisms across diverse environments.

The distribution of QR-ACS applications across domains is summarized in Fig. 8. The figure illustrates the distribution of QR-ACS applications across different operational domains. Home security and university laboratory environments constitute the largest proportion of reported implementations, reflecting the technology's suitability for cost-sensitive and semi-controlled environments. Public facilities and IoT-integrated smart systems also represent significant adoption sectors. In contrast, high-security and healthcare applications show comparatively lower representation, indicating potential research and deployment gaps in environments requiring stricter compliance, scalability validation, and regulatory assurance.

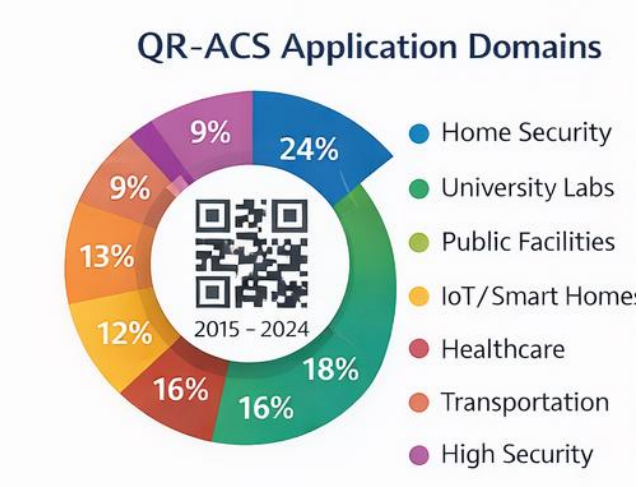


Fig. 8. Distribution of QR-ACS application domains.

Overall, these recurring challenges indicate the need for continued research and refinement of QR-ACS to improve reliability, scalability, and practical deployment.

The ongoing development of QR code technologies and their integration with emerging security frameworks suggest that QR-ACS remains an active area for both academic investigation and real-world application. Fig. 9 presents a structured taxonomy of QR Code-Based Access Control Systems (QR-ACS), synthesizing the diverse implementations identified in the reviewed literature. The taxonomy categorizes QR-ACS into five primary classes based on authentication architecture and security mechanisms:

- Static QR Systems, which rely on fixed credentials and minimal computational complexity, are predominantly observed in early home security and institutional prototypes (Tables I and III). While cost-effective, these systems remain vulnerable to duplication and replay attacks.
- Dynamic QR Systems, including time-based and session-based implementations, are introduced to mitigate replay and reuse risks. These systems are commonly applied in university laboratories, transportation ticketing, and public facility management (Tables III and X), where temporal validity enhances security robustness.
- Encrypted QR Systems, incorporate cryptographic algorithms such as AES, RSA, or hybrid encryption to protect embedded credentials. As shown in Tables III, IV, V, and XI, encryption-based approaches are widely adopted in healthcare, high-security, and emergency-response environments to address confidentiality and integrity concerns.
- Multi-Factor QR Systems, combining QR authentication with biometric verification, one-time passwords (OTP), or blockchain logging, are increasingly reported in office, financial, and high-security settings (Tables II, V, and VIII). These systems significantly strengthen authentication assurance but introduce higher computational and infrastructural complexity.
- IoT-Integrated QR Systems, observed across smart homes, industrial IoT, and facility management contexts (Tables VI, VII, and IX), embed QR authentication within connected infrastructures, enabling real-time monitoring and context-aware access control.

This taxonomy provides a conceptual framework that clarifies architectural distinctions, highlights evolutionary security enhancements, and supports comparative analysis across application domains. By structuring the literature into these categories, the taxonomy contributes to a clearer understanding of design trends and emerging directions in QR-ACS research. Fig. 9 presents a structured taxonomy of QR Code-Based Access Control Systems (QR-ACS).

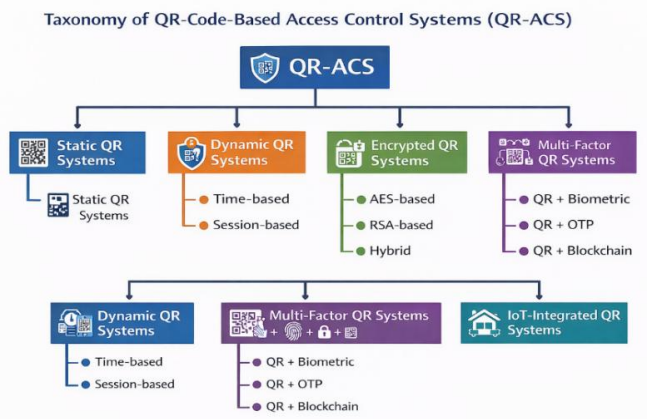


Fig. 9. Two-dimensional architectural model of QR-ACS.

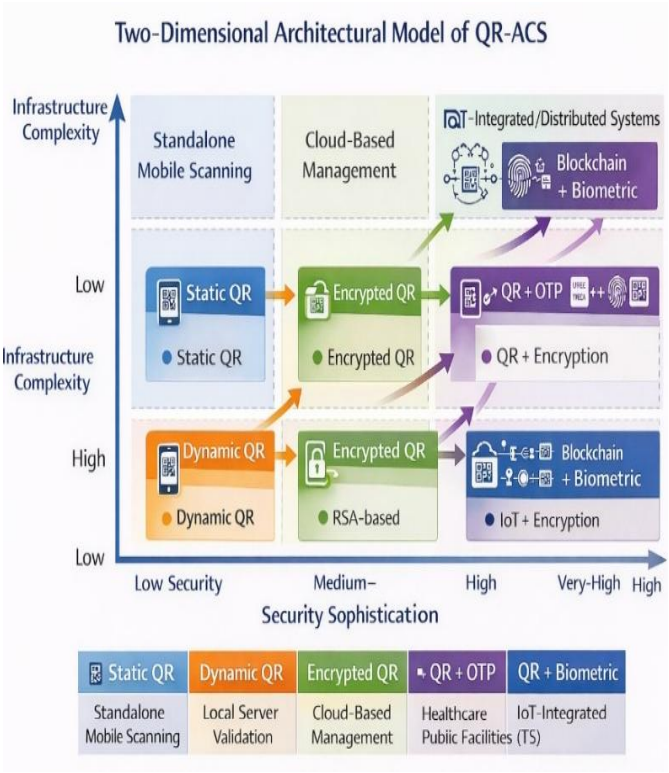


Fig. 10. Two-dimensional architectural model of QR-ACS.

Fig. 10 extends the taxonomy into a two-dimensional architectural model that classifies QR-ACS implementations based on security sophistication and infrastructure complexity. Systems positioned in the lower-left quadrant, such as static QR implementations, exhibit low computational overhead but limited resistance to duplication and replay attacks. Moving toward the upper-right quadrant, multi-factor and blockchain-integrated systems demonstrate enhanced security robustness at the cost of increased infrastructural and computational demands. This model highlights the inherent trade-off between deployment simplicity and security strength, enabling clearer benchmarking of QR-ACS designs across different operational environments.

Table XIII provides a structured comparison of QR-ACS categories derived from the proposed taxonomy. The analysis shows that as systems evolve from static to multi-factor and blockchain-integrated architectures, replay resistance and authentication assurance increase significantly. However, this improvement is accompanied by greater infrastructural requirements and system complexity. The table reinforces the trade-offs illustrated in the two-dimensional architectural model and provides a practical reference for selecting QR-ACS architectures based on deployment constraints and security requirements.

V. Current Technologies and Applications

The application of QR Code-Based Access Control Systems (QR-ACS) extends across a wide range of environments, where the technology is employed to support security requirements and improve operational processes. This section reviews the key technologies currently associated with QR-ACS and discusses how they are applied in real-world contexts across different sectors.

Fig. 11 illustrates several application environments and technological trends that have influenced the deployment of QR code-based access control systems. These environments include digital services, smart environments, and data-driven infrastructures where secure and contactless authentication methods are increasingly required. The figure also highlights broader technological developments such as remote monitoring, data integration, and privacy-aware systems, which have contributed to expanding the role of QR-based technologies in modern security and operational management.

TABLE XIII. Comparative Characteristics of QR-ACS Categories

Category	Credential Type	Replay Resistance	Infrastructure Need	Typical Domains (Tables)
Static QR	Fixed	Low	Low	Home Security (T1), Early Labs (T3)
Dynamic QR	Time/Session-based	Medium–High	Medium	Labs (T3), Transportation (T10)
Encrypted QR	AES/RSA	High	Medium	Healthcare (T11), Public Facilities (T4), High Security (T5)
Multi-Factor QR	QR + OTP/Bio	Very High	High	Offices (T2), Finance (T8), High Security (T5)
IoT-Integrated QR	Context-aware	Variable	High	Smart Homes (T6), Industrial IoT (T7), Smart Buildings (T9)

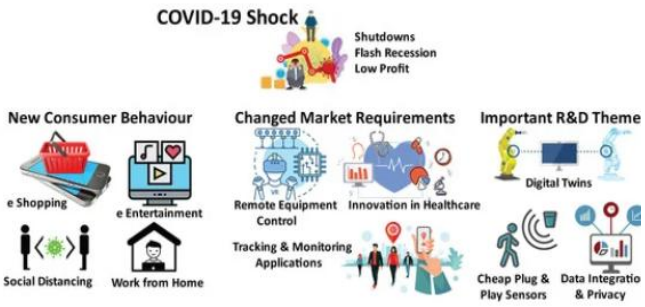


Fig. 11. Overview of application environments for QR code-based access control systems, including smart homes, healthcare, industrial settings, and high-security facilities [63].

A. Technological Innovations

Recent developments in QR-ACS have focused on improving the security of data encoded within QR codes through the use of cryptographic techniques. Many proposed systems employ encryption algorithms such as AES (Advanced Encryption Standard) and RSA (Rivest–Shamir–Adleman) to protect sensitive information from unauthorized access or manipulation. Previous studies note that QR codes have increasingly been adopted for security-related applications when combined with encryption mechanisms that support data confidentiality and integrity [64].

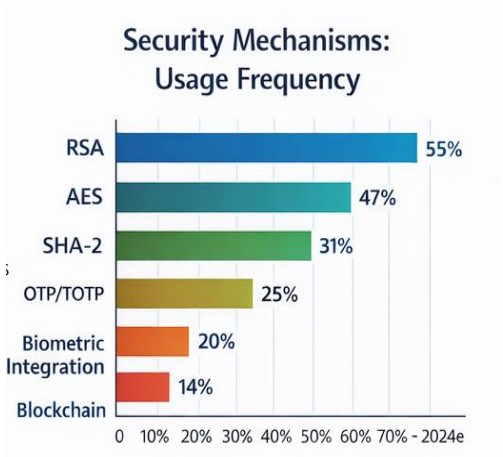


Fig. 12. Security mechanism usage frequency in QR-ACS.

In addition, the integration of cloud-based technologies has enabled more flexible management of QR codes. Cloud platforms allow QR codes to be updated dynamically and managed centrally, reducing the need for physical interaction and supporting scalable deployment in distributed environments [65]. The relative adoption frequency of security mechanisms is illustrated in Fig. 12. The figure compares the relative frequency of security mechanisms integrated within QR-ACS implementations. As shown, RSA and AES encryption dominate the literature, reflecting their widespread adoption for protecting QR-encoded credentials. Hash-based mechanisms such as SHA-2 are moderately represented, primarily in login authentication contexts. Multi-factor authentication techniques, including OTP/TOTP and biometric integration, appear less frequently but demonstrate increasing adoption in recent years. Blockchain-based logging mechanisms remain emerging but show promise for enhancing transparency and tamper resistance. This distribution suggests that encryption remains foundational, while decentralized and biometric security layers are still developing.

B. Biometric Integration

Another area of development in QR-ACS involves the combination of QR code authentication with biometric verification methods, such as facial recognition and fingerprint scanning. In these systems, QR codes are typically used as an initial authentication step, followed by biometric verification to confirm user identity. Studies suggest that combining encrypted QR data with biometric information can provide an additional layer of security compared to traditional single-factor access control methods [66].

This approach has been explored particularly in environments that require higher security levels, where layered authentication mechanisms can help reduce the risk of unauthorized access.

C. IoT and Smart Environments

In smart homes and smart buildings, QR-ACS are often integrated with Internet of Things (IoT) infrastructures to support more adaptive access control solutions. These systems can interact with sensors, actuators, and monitoring devices to adjust access permissions based on contextual factors such as time, location, or environmental conditions. Existing research indicates that IoT-enabled QR-based access control frameworks can enhance system responsiveness and support real-time monitoring when combined with encrypted communication and authentication mechanisms [67].

For example, access to specific areas within a smart building may be regulated not only by user identity but also by predefined environmental or operational conditions.

D. Public and Commercial Applications

QR-ACS has also been adopted in a variety of public and commercial settings, including transportation systems, entertainment venues, and corporate facilities. In such environments, QR codes are commonly used to manage access efficiently, particularly in high-traffic scenarios. Prior studies reported that QR-based access control can support crowd management and streamline entry procedures by reducing reliance on physical tickets or access cards [67].

In public transportation systems, for instance, QR codes are frequently used to facilitate contactless entry and verification, contributing to improved operational efficiency during peak usage periods.

E. Healthcare and Institutional Use

In healthcare and institutional environments, QR-ACS are primarily employed to control access to restricted areas and sensitive information, such as patient records and clinical facilities. These systems can be configured to grant access only to authorized personnel while maintaining detailed access logs for auditing and compliance purposes. Research has shown that QR-based authentication mechanisms, particularly when combined with biometric verification, can support enhanced protection of patient data and restricted zones [68].

During health emergencies, including the COVID-19 pandemic, QR-based systems have also been utilized to manage patient flow and visitor access in a contactless manner, helping institutions maintain safety protocols while reducing physical interaction [65].

VI. Discussion

The review of QR Code-Based Access Control Systems (QR-ACS) provides several insights into both the strengths of these systems and the challenges that continue to affect their deployment, as in Table XIV. By examining studies across different application domains, this review highlights recurring technical, operational, and organizational issues that influence the effectiveness of QR-ACS, while also pointing to areas where further research may contribute to more robust implementations.

TABLE XIV. Security Strength Comparison

Mechanism	Duplication Resistance	Replay Protection	Scalability	Complexity
Static QR	Low	Low	High	Low
Encrypted QR	Medium	Medium	Medium	Medium
QR + OTP	High	High	Medium	Medium
QR + Biometric	Very High	High	Low-Medium	High
Blockchain QR	Very High	Very High	Medium	Very High

To enhance the practical applicability of the proposed taxonomy, a qualitative risk assessment matrix is introduced. The matrix evaluates key threats based on likelihood and impact, enabling prioritization of security measures, as Table XV. The analysis indicates that duplication and replay attacks represent the most critical risks in QR-ACS, particularly in static implementations, while biometric spoofing, although less frequent, poses severe consequences in high-security environments.

TABLE XV. Risk Assessment of QR-ACS Threats

Threat Type	Likelihood	Impact	Risk Level
QR Code Duplication	High	Medium	High
Replay Attacks	Medium	High	High
Biometric Spoofing	Low	Very High	High
Device Theft	Medium	Medium	Medium
Insider Misuse	Low	High	Medium

The findings indicate a clear trade-off between system security and deployment complexity. While multi-factor and blockchain-based QR-ACS provide higher resistance against attacks, they require significant infrastructural support and computational overhead. Conversely, lightweight QR systems offer scalability but remain vulnerable to duplication and replay attacks. This trade-off suggests that no single architecture is universally optimal; instead, system design should be guided by application-specific security requirements and operational constraints.

Moreover, the increasing integration of QR-ACS with IoT environments introduces new attack surfaces, particularly related to device-level vulnerabilities and network security. This highlights the need for cross-layer security frameworks that combine cryptographic protection, device authentication, and real-time monitoring.

VII. Challenges in Current QR-ACS

One of the most frequently discussed challenges in the literature relates to security vulnerabilities. Although many systems incorporate encryption techniques and, in some cases, biometric verification, QR-ACS can still be exposed to risks such as code duplication, spoofing, and unauthorized access. In particular, systems that rely on static QR codes may be more susceptible to misuse unless supported by dynamic generation and validation mechanisms.

Another limitation concerns the dependence on specific hardware, especially smartphones and dedicated scanning devices. While this reliance is feasible in many contexts, it can restrict deployment in environments where access to such technology is limited or economically impractical. To provide a structured understanding of these risks, Fig. 13 illustrates the threat landscape associated with QR-ACS implementations. Fig. 6 illustrates the threat landscape associated with QR Code-Based Access Control Systems (QR-ACS). The model categorizes vulnerabilities into code-level threats (e.g., duplication and replay attacks), authentication-layer threats (e.g., biometric spoofing and static QR reuse), and infrastructure-level risks (e.g., insider misuse and device theft). The figure highlights that static QR implementations are particularly vulnerable to duplication and replay-based exploitation, whereas systems integrating biometric verification and dynamic QR generation exhibit reduced exposure but introduce new attack surfaces related to biometric spoofing and device compromise. This layered threat perspective reinforces the need for multi-factor authentication and dynamic credential management in next-generation QR-ACS frameworks.

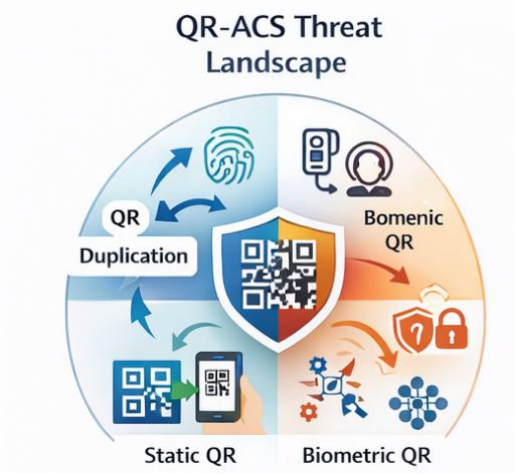


Fig. 13. QR-ACS threat model landscape.

Privacy considerations also emerge as a significant concern, especially in systems that integrate biometric data or personal identifiers. Ensuring compliance with data protection regulations, such as the General Data Protection Regulation (GDPR), is essential for maintaining user trust and safeguarding sensitive information.

In addition, the lack of standardized protocols across different QR-ACS implementations often leads to interoperability challenges. As these systems are adopted in diverse sectors, inconsistencies in design and communication standards can hinder seamless integration with existing security infrastructures.

VIII. Future Research Directions

The findings of this review suggest several directions for future research. Enhancing security protocols remains a priority, particularly through the development of cryptographic techniques that can be efficiently embedded within QR-based systems. Emerging approaches, such as blockchain-based access logging, have also been discussed in the literature as potential solutions for improving transparency and traceability. Expanding accessibility represents another important research avenue. Future work may explore alternatives to smartphone-dependent systems, including low-cost scanning devices or hybrid authentication models that combine QR codes with other access control mechanisms.

To highlight areas requiring further investigation, Fig. 14 presents a structured research gap matrix. The figure presents a structured research gap matrix synthesizing the maturity of QR-ACS research across key dimensions. While encryption techniques are widely implemented across most domains, blockchain integration and large-scale real-world validation remain underexplored, particularly in public facilities and high-security environments. Scalability testing is inconsistently addressed, with many studies limited to prototype-level validation. The matrix highlights that although technical feasibility has been demonstrated, empirical deployment studies and interoperability frameworks are still insufficient. These gaps underscore the need for standardized evaluation methodologies and cross-domain benchmarking in future QR-ACS research.

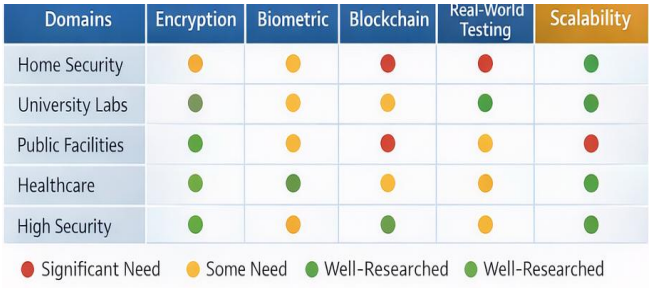


Fig. 14. Research gap matrix (2015–2024).

Dynamic QR code generation is also widely recognized as a promising approach to strengthening security. Techniques that enable real-time code generation and expiration could reduce the risk of unauthorized reuse, while adaptive methods based on usage patterns may further improve access control reliability.

Privacy-preserving frameworks require continued attention as well. Research focusing on data anonymization, secure storage, and regulatory compliance can contribute to the development of QR-ACS that balances security requirements with user privacy.

Finally, cross-industry standardization is essential for supporting large-scale deployment. Collaborative efforts among academic researchers, industry stakeholders, and regulatory authorities may facilitate the creation of interoperable standards that enhance scalability and adoption across different application domains.

IX. Conclusion

This systematic review has examined the evolution and current state of QR Code-Based Access Control Systems (QR-ACS), drawing on a wide range of studies that reflect their application across residential, institutional, commercial, and high-security environments. The reviewed literature indicates that QR-ACS offers flexible and cost-effective solutions for access control, particularly when combined with encryption techniques, biometric verification, and cloud-based management.

At the same time, the findings highlight several persistent challenges that must be addressed to ensure reliable and secure deployment. These challenges include security vulnerabilities, privacy concerns, hardware dependency, and interoperability limitations. Addressing such issues is critical for improving the robustness of QR-ACS and supporting their integration into complex security infrastructures.

Future research plays a key role in advancing this field. Continued efforts are needed to develop stronger security mechanisms, explore more inclusive and accessible system designs, and establish standardized frameworks that promote interoperability and scalability. Equally important is the need to incorporate privacy-preserving approaches that align with international data protection regulations and user expectations.

Overall, QR Code-Based Access Control Systems demonstrate significant potential to influence the future of access management in increasingly digital environments. With sustained research, careful system design, and cooperation between academia, industry, and regulatory bodies, QR-ACS can evolve into more secure, efficient, and user-centered access control solutions.

References

[1] M. Adedoyin and F. Olukoya, Development of a Smart Lock System using QR Code Technology. AJERD. doi: 10.53982/ajerd.

[2] G. Tejaswai et al., "Development of an IoT-Based QR Code Access Control and Payment System using Arduino and ESP8266"," Journal of Science and Technology, vol. 9, no. 06, pp. 20–32, 2024, doi: 10.46243/jst.2024.v9.i06.pp20-32.

[3] G. Petrea, C. Vlad, C. Aramă, M. Crăciun, and B. Ionescu, "QR code based access control system for hotels," Analele Universităţii" Dunărea de Jos" din Galaţi. Fascicula II, Matematică, fizică, mecanică teoretică/Annals of the" Dunarea de Jos" University of Galati. Fascicle II, Mathematics, Physics, Theoretical Mechanics, vol. 43, pp. 121–126, 2020.

[4] J. B. Awotunde, A. E. Adeniyi, A. L. Imoize, Y. Mejdoub, and Z. Abdualazizu, "A Mobile Visitor Management System Using a QR Code and PIN for Access Control," in International Conference on Connected Objects and Artificial Intelligence, Springer, 2024, pp. 192–198.

[5] S. Kaushik, "Steganography with Visual Secret Sharing Scheme," International Journal of Advanced Computer Science and Applications, vol. 2, Nov. 2011.

[6] M. Hara, "Development and popularization of QR code—Code development pursuing reading performance and market forming by open strategy—," National Institute of Advanced Industrial Science and Technology, 2019.

[7] P. C. Huang, C. C. Chang, Y. H. Li, and Y. Liu, "Efficient access control system based on aesthetic QR code," Pers. Ubiquitous Comput., vol. 22, no. 1, pp. 81–91, Feb. 2018, doi: 10.1007/s00779-017-1089-y.

[8] A. Jain, A. Panwar, Mohd. Azam, and R. Khanam, "Smart door access control system based on QR code," International Journal of Informatics and Communication Technology (IJ-ICT), vol. 12, no. 2, p. 171, Aug. 2023, doi: 10.11591/ijict.v12i2.pp171-179.

[9] R. Khatri, T. Shaikh, N. Gundecha, O. Kathe, K. Tiwari, and B. Tech, "QR Code Based Access Control System," 2023. Accessed: Nov. 17, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:259905457.

[10] T. L. Norman, Electronic access control. Elsevier, 2011.

[11] M. R. Rieback, B. Crispo, and A. S. Tanenbaum, "The evolution of RFID security," IEEE Pervasive Comput., vol. 5, pp. 62–69, 2006.

[12] E. Edwards and P. Orukpe, "Development of a RFID Based Library Management System and User Access Control," Nigerian Journal of Technology, vol. 33, no. 4, p. 574, Sep. 2014, doi: 10.4314/njt.v33i4.19.

[13] J. Yang, Y. Zhang, and C. J. M. Lanting, "Exploring the Impact of QR Codes in Authentication Protection: A Study Based on PMT and TPB," Wirel. Pers. Commun., vol. 96, no. 4, pp. 5315–5334, Oct. 2017, doi: 10.1007/s11277-016-3743-5.

[14] G. de Seta, "QR code: The global making of an infrastructural gateway," Global Media and China, vol. 8, no. 3, pp. 362–380, Sep. 2023, doi: 10.1177/20594364231183618.

[15] S. Kumar and M. L. Nirmal, "Advanced Visual Cryptography Secret Sharing Schemes Based on QR Codes: A Survey," Dogo Rangsang Research Journal, vol. 11, no. 01, 2021.

[16] A. AWASTHI and A. PANDEY, "Steganography with Visual Secret Sharing Scheme Based QR Code Application: A Survey," Dogo Rangsang Research Journal, vol. 10, no. 07, Jul. 2020.

[17] U. Emenike, "Design and Implementation of a Business Networking App Using or Code," 2016, doi: 10.13140/RG.2.2.34012.60802.

[18] H. Gupta, S. Avasthi, and D. a, "ScanKaro: An QR Code-Based Menu Application for Restaurants," in Artificial Intelligence and Communication Technologies, Soft Computing Research Society, 2023, pp. 945–951. doi: 10.52458/978-81-955020-5-9-90.

[19] M. Kamal Hossain, S. Software Engineer, and M. Mynuddin, "International Journal of Modern Embedded System (IJMES) Design and Implementation of Smart Home Security System Prodip Biswas 2 nd," 2014. [Online]. Available: http://ijmes.info

[20] S. Dharme, "Door Lock Security System Using Recent Technology," Int. J. Res. Appl. Sci. Eng. Technol., vol. 10, no. 1, pp. 684–686, Jan. 2022, doi: 10.22214/ijraset.2022.39812.

[21] N. Susmitha, P. Sindhuja, R. M. Krishna, and B. Mamatha, "HOME SECURITY AND AUTOMATION SYSTEMBASED ON AI," Complexity International Journal (CIJ), vol. 25, pp. 5–6, 2021, [Online]. Available: http://cij.org.in/Currentvolumissue2502.aspx1593

[22] K. T. Reddy, "Intelligent Door Lock System with Face Recognition," Int. J. Res. Appl. Sci. Eng. Technol., vol. 8, no. 5, pp. 364–371, May 2020, doi: 10.22214/ijraset.2020.5060

[23] N. Ghodekar, K. Wagh, R. Kale, S. Kadekar, M. Vijay, and N. Kukre, "Door Lock Control using Wireless Biometric," International Research Journal of Engineering and Technology, 2020, [Online]. Available: www.irjet.net

[24] S. Lakshminarayanan et al., "Home Security System Based on Finger Print and Face Recognition," International Research Journal of Engineering and Technology, 2020.

[25] S. Marrapu, S. Satyanarayana, V. ArunKumar, and J. D. S. K. Teja, "Smart home based security system for door access control using smart phone," International Journal of Engineering and Technology(UAE), vol. 7, no. 1, pp. 249–251, 2018, doi: 10.14419/ijet.v7i1.9247.

[26] P. Y. Lin, W. S. Lan, Y. H. Chen, and W. C. Wu, "A Confidential QR Code Approach with Higher Information Privacy," Entropy, vol. 24, no. 2, Feb. 2022, doi: 10.3390/e24020284.

[27] M. F. Jamal and N. Rahim, "Fingerprint and Qr Code Based Authentication System at Pusat Servis Komputer JB," Applied Information Technology And Computer Science, vol. 4, no. 1, pp. 205–223, 2023, doi: 10.30880/aitcs.2023.04.01.013.

[28] P. Satanasaowapak, W. Kawseewai, S. Promlee, and A. Vilamat, "Residential access control system using QR code and the IoT," International Journal of Electrical and Computer Engineering, vol. 11, no. 4, pp. 3267–3274, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3267-3274.

[29] V. Susukailo and Y. Lakh, Access control system based on encryption in QR-Code technology. IEEE, 2018. doi: 10.1109/IDAACS-SWS.2018.8525779.

[30] A. F. M. Fauzi, N. N. Mohamed, H. Hashim, and M. A. Saleh, Development of Web-Based Smart Security Door Using QR Code System. IEEE, 2020.

[31] I. Ishrat, W. M. Ali, S. Ghani, S. Sami, M. Waqas, and F. Aftab, "SMART DOOR LOCK SYSTEM WITH AUTOMATION AND SECURITY," Sci.Int.(Lahore), vol. 29, no. 1, pp. 73–76, 2017.

[32] H. Alsalem et al., "Laboratory Access Implementing QR Code Authentication Using OTP," International Journal on Cybernetics & Informatics, vol. 12, no. 5, pp. 121–141, Aug. 2023, doi: 10.5121/ijci.2023.120511.

[33] H. Park, T. Kim, G. Kim, W. Jang, K. Lee, and S. Y. Lee, "Improvement of QR Code Access Control System Based on Lamport Hash Chain," in Advances in Intelligent Systems and Computing, Springer Verlag, 2020, pp. 824–833. doi: 10.1007/978-3-030-22263-5_79.

[34] L. Kakwete Musambo and J. Phiri, "Student Facial Authentication Model based on OpenCV's Object Detection Method and QR Code for Zambian Higher Institutions of Learning," 2018.

[35] E. A. Amusan, A. O. Popoola, and S. A. O. Ogirima, "Securing Logins in Electronic Examination Systems for Tertiary Institutions Using Quick Response Code (QR) Technology and Multiple Hashing Algorithms," Asian Journal of Research in Computer Science, pp. 23–34, Sep. 2021, doi: 10.9734/ajrcos/2021/v11i330264.

[36] S. Belguith, S. P. Gochhayat, M. Conti, and G. Russello, "Emergency access control management via attribute based encrypted QR codes," in 2018 IEEE Conference on Communications and Network Security, CNS 2018, Institute of Electrical and Electronics Engineers Inc., Aug. 2018. doi: 10.1109/CNS.2018.8433186.

[37] Y. Hong, "Design of Intelligent Access Control System Based on des Encrypted QR Code," in Proceedings of 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications, AEECA 2020, Institute of Electrical and Electronics Engineers Inc., Aug. 2020, pp. 1005–1008. doi: 10.1109/AEECA49918.2020.9213475.

[38] X. Xiong, S. Zhang, S. Wang, and J. Crawford, "Intelligent Identity Authorization System Design Based on WSN and Cloud Platform," U.P.B. Sci. Bull., Series C, vol. 82, no. 3, p. 2020, 2020.

[39] Y. Yan, Z. Zou, H. Xie, Y. Gao, and L. Zheng, "An IoT-Based Anti-Counterfeiting System Using Visual Features on QR Code," IEEE Internet Things J., vol. 8, no. 8, pp. 6789–6799, Apr. 2021, doi: 10.1109/JIOT.2020.3035697.

[40] J. Y. Kim, Y. Hoon Jung, D. S. Yang, and M.-S. Jun, "Implementation of Integrated Authentication Service using Blockchain and One Time QR Code for Access Control in U-city Environment," International Journal of Advanced Research in Big Data Management System, vol. 3, no. 2, pp. 15–20, 2019, doi: 10.21742/ijarbms.2019.3.2.03.

[41] Z. H. Choudhury and M. M. A. Rabbani, "Biometric Passport for National Security Using Multibiometrics and Encrypted Biometric Data Encoded in the QR Code," Journal of Applied Security Research, vol. 15, no. 2, pp. 199–229, Apr. 2019, doi: 10.1080/19361610.2019.1630226.

[42] Y. Chung, S. Jung, J. Kim, J. Lee, and J. Cha, "LED-QR Authentication Technology for Access Control and Security," International journal of advanced smart convergence, vol. 4, no. 2, pp. 69–75, Nov. 2015, doi: 10.7236/ijasc.2015.4.2.69.

[43] A. Mohammed Ali and A. K. Farhan, "Enhancement of QR Code Capacity by Encrypted Lossless Compression Technology for Verification of Secure E-Document," IEEE Access, vol. 8, pp. 27448–27458, 2020, doi: 10.1109/ACCESS.2020.2971779.

[44] A. Palaniappan, L. P. Veilumuthu, and R. P. S. Louis, "Enhancing Security Through QR Code and Enriched Blowfish Cryptography for Sensitive Data," Springer, Cham, 2024.

[45] K. A. Hashim, H. H. Qasim, A. E. Hamzah, O. A. Hasan, and M. Al-Jadiri, "Door lock system based on internet of things and Bluetooth by

using Raspberry Pi," Bulletin of Electrical Engineering and Informatics, vol. 12, no. 5, pp. 2753–2762, Oct. 2023, doi: 10.11591/eei.v12i5.5134.

[46] F. Yahya, F. A. Hazri, and M. S. M. Kassim, "Entrance monitoring and authentication IoT based system," in AIP Conference Proceedings, American Institute of Physics Inc., Nov. 2022. doi: 10.1063/5.0119671.

[47] A. Fadiga, M. Agarwal, and P. Kaushik, "Home security based on Internet of things Review paper," International Research Journal of Engineering and Technology, 2022.

[48] T. Adiono, S. Fuada, S. Feranti Anindya, I. G. Purwanda, and M. Y. Fathany, "IoT-Enabled Door Lock System," 2019.

[49] K. N. L, B. G. Kumar, D. V D, H. G. Kumar R, and A. Professor, "Home Automation System with Security using Raspberry-Pi," International Research Journal of Engineering and Technology, 2020.

[50] Atanas Plamenov Karaguiozov, "QReact A Generic Framework to Create and Use QR Codes and a Usage Case in the Field of Access Control under Android."

[51] A. M. Al-Ghaili, H. Kasim, M. Othman, and W. Hashim, "QR code based authentication method for IoT applications using three security layers," Telkomnika (Telecommunication Computing Electronics and Control), vol. 18, no. 4, pp. 2004–2011, 2020, doi: 10.12928/TELKOMNIKA.V18I4.14748.

[52] K. C. Liao and W. H. Lee, "A novel user authentication scheme based on QR-code," Journal of Networks, vol. 5, no. 8, pp. 937–941, 2010, doi: 10.4304/jnw.5.8.937-941.

[53] G. Ali, M. A. Dida, and A. E. Sam, "A secure and efficient multi‐factor authentication algorithm for mobile money applications," Future Internet, vol. 13, no. 12, Dec. 2021, doi: 10.3390/fi13120299.

[54] Y. Wahyu Agung Prasetyo, R. Rahim, M. A. Manuhutu, and S. Sujito, "QRIS and GOST: A Symbiotic Approach for Secure QR Code Transactions," SSRG International Journal of Electronics and Communication Engineering, vol. 10, no. 5, pp. 138–147, May 2023, doi: 10.14445/23488549/IJECE-V10I5P113.

[55] Y. Sil Lee, N. Hyun Kim, H. Lim, H. Jo, and H. Jae Lee, "Online Banking Authentication System using Mobile-OTP with QR-code," 2010.

[56] J. Sa-Ngiampak et al., "LockerSwarm: An IoT-based Smart Locker System with Access Sharing," 2019.

[57] B. Muthukumar, J. Albert Mayan, G. Nambiar, and D. Nair, "QR Code and Biometric Based Authentication System for Trains," in IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, Oct. 2019. doi: 10.1088/1757-899X/590/1/012010.

[58] D. Conde-Lagoa, E. Costa-Montenegro, F. J. González-Castaño, and F. Gil-Castiñeira, Secure eTickets Based on QR-Codes with User-Encrypted Content. IEEE, 2010.

[59] A. M. S. Pangan, I. L. Lacuesta, R. C. Mabborang, and F. P. Ferrer, "Authenticating Data Transfer Using RSA-Generated QR Codes," pp. 18–38, 2022.

[60] M. E. Sankar, T. Sai Divya, and R. Gunasri, "Delegated Authorization Framework for the Services Using or Code Generation," 2020.

[61] B. Olaiya and O. Onyekachi, "A QR-Code-Based Identity Management System for Monitoring Hostel Residents," Mountain Top University, 2021.

[62] Y. Rahayu, L. Afif, and P. J. Soh, "Design and development of smart lock system based QR-Code for library's locker at Faculty of Engineering, Universitas Riau," SINERGI, vol. 26, no. 3, p. 379, Oct. 2022, doi: 10.22441/sinergi.2022.3.013.

[63] M. Umair, M. A. Cheema, O. Cheema, H. Li, and H. Lu, "Impact of COVID-19 on IoT Adoption in Healthcare, Smart Homes, Smart Buildings, Smart Cities, Transportation and Industrial IoT," Sensors, vol. 21, no. 11, p. 3838, Jun. 2021, doi: 10.3390/s21113838.

[64] K. Saranya, R. S. Reminaa, and S. Subhitsha, Modern applications of QR-Code for security. 2016.

[65] S. Suman Rajest Professor, R. Regin Assistant Professor, and S. R. Assistant Professor, "International Journal of Human Computing Studies (IJHCS) A QR Code-Based Real-Time Auditing System for Safe Online Data Storage," 2024. [Online]. Available: https://journals.researchparks.org/index.php/IJHCS

[66] Z. H. Choudhury and M. M. A. Rabbani, "Biometric passport for National Security Using Multibiometrics and encrypted biometric data encoded in the QR code," Journal of Applied Security Research, pp. 199–229, 2020.

[67] A. Y. Chaudhari, J. Kulkarni, U. Dube, A. Ghorpade, S. Sakore, and A. K. Gupta, A Biometric Authentication System By Embedding Biometrics in QR Codes. ieee, 2024.

[68] D. Reddy Rachapalli and H. Kumar Kalluri, "Disseminating the Authentication Process Based on Secure RGVSS Multi-Biometric Template Encryption through QR Code in Health Care Informatics," International Journal on Emerging Technologies, vol. 10, no. 3, pp. 370–378, 2019.

Real-Traffic-Trained Intelligent IDS for Advanced Cyberattack Detection in Enterprise Networks

Dalila Naira Chinchay, Rodrigo Calderón Ari, Liset S. Rodriguez-Baca
Professional School of System Engineering, Cesar Vallejo University, Lima, Peru

***Abstract*—Early detection of cyberattacks remains a major challenge in enterprise networks due to encrypted traffic, protocol diversity, and highly dynamic service behavior. This study evaluates a machine learning-based intrusion detection system trained on real enterprise traffic captured over 20 working days under operational conditions. A total of 1,163,014 packets were collected and complemented with controlled attack traffic, including DDoS, brute force, botnet, SQL injection, port scanning, privilege escalation, and service exploitation scenarios. After flow-based feature extraction and preprocessing, six supervised learning models were evaluated under the same data partition and validation settings. Among them, Random Forest achieved the best overall performance, with precision, recall, and F1-score above 0.999 and an AUC of 0.9994 on the collected dataset. These findings suggest that training with real traffic can improve IDS performance under realistic enterprise conditions. However, further validation across additional organizations and time periods is required to confirm generalizability.**

***Keywords*—IDS; machine learning; cyberattacks; cybersecurity; intrusion detection**

I. Introduction

The growing reliance of organizations on digital infrastructures has intensified their exposure to increasingly sophisticated and hard-to-anticipate cyberattacks. In recent years, threats based on evasion techniques, zero-day vulnerability exploitation, and malicious automation have proven capable of easily bypassing traditional security systems. These challenges have driven the evolution of Intrusion Detection Systems (IDS), which aim to identify anomalous behaviors or attack-related patterns with ever-higher levels of accuracy.

Recent literature agrees that approaches based on artificial intelligence and machine learning offer significant advantages for early intrusion detection, particularly when dealing with unknown attacks or rapidly evolving variants. Studies such as that of Shamshair et al. [1] highlight that AI-based models can adapt to emerging threats with a level of generalization superior to mechanisms relying solely on signature-based detection. Similarly, applied research in IoT environments and enterprise networks shows that advanced techniques based on deep learning and graph-based methods have achieved improved detection of complex malicious behaviors [2], [3].

Despite these advances, a key limitation remains: most existing studies rely on synthetic or historical datasets that do not adequately represent the real behavior of modern networks. Widely used datasets such as KDD99, NSL-KDD, or UNSW-NB15 exhibit artificial structures, obsolete protocols, or traffic volumes that differ substantially from real traffic captured in an operational organization. As a result, models that achieve high performance in controlled environments often lose effectiveness when deployed in networks dominated by encrypted protocols, highly variable flows, and distributed application workloads.

Recent systematic reviews confirm this gap. Rehman et al. [4] emphasize that machine learning–based IDS requires datasets that are representative of current network traffic, while Chen et al. [5] underline that model performance largely depends on the quality of the data used during training. Although recent studies have reported promising IDS results using machine learning and deep learning, many of them rely on public benchmark datasets or controlled laboratory environments. These settings often underrepresent encrypted enterprise traffic, fluctuating service workloads, and the coexistence of modern protocols such as QUIC and TLS/HTTPS. As a result, it remains unclear to what extent the reported performance of these models can be sustained under real operational conditions.

The main contributions of this study are threefold: 1) the construction of a hybrid dataset that combines real enterprise traffic with controlled attack scenarios; 2) the comparative evaluation of six supervised machine learning models under identical preprocessing, partitioning, and validation settings; and 3) empirical evidence that real-traffic-based training can support highly accurate intrusion detection under realistic enterprise network conditions.

II. Related Work

Intrusion detection has been widely studied within the field of cybersecurity, and its evolution has been marked by the progressive incorporation of machine learning techniques. Over the past decade, numerous studies have compared traditional algorithms such as decision trees, support vector machines, and k-nearest neighbors with more recent models based on ensemble methods or deep learning, demonstrating significant improvements in accuracy, predictive capability, and the detection of hidden patterns.

Shamshair et al. [1] conducted a comparative analysis of various AI-based methods applied to zero-day attack detection, demonstrating that ensemble-based models such as Random Forest and Gradient Boosting exhibit greater stability under traffic variations and achieve better performance in real-world scenarios. In IoT environments, where network behavior tends to be more dynamic and unpredictable, Villegas et al. [2] introduced an innovative approach based on dynamic graphs and neural networks, achieving high accuracy in the identification of

complex attacks. Complementarily, Villegas et al. [3] evaluated the effectiveness of an adaptive deep learning–based IDS, reporting substantial improvements against multi-stage attacks and low-frequency malicious patterns.

A systematic review conducted by Rehman et al. [4] highlights the advantages of machine learning techniques and their applications in IDS, reinforcing arguments related to current trends and existing research gaps. In line with this perspective, Ferrag et al. [6] emphasize that training with real data strengthens the model's generalization capability and reduces the occurrence of false positives, which remain one of the main challenges of traditional IDS. Similarly, Chen et al. [5] propose a hybrid NIDS + HIDS system that improves detection performance by combining network- and host-based features, representing a modern approach to machine learning–based IDS.

Studies focused specifically on intrusion detection have also examined the comparative behavior of supervised classifiers such as Random Forest, Decision Tree, KNN, AdaBoost, and XGBoost under different traffic conditions. In general, ensemble-based models tend to provide more stable results than single classifiers when the data exhibit variability, noise, or class imbalance.

Likewise, Jijo et al. [7] provide a comprehensive review of the decision tree algorithm, highlighting its importance as a foundation for advanced models such as Random Forest and Extra Trees. In the context of critical infrastructures, Majidi et al. [8] applied ensemble techniques and autoencoders for intrusion detection in power grid networks, demonstrating their robustness against false data injection attacks.

From an algorithmic perspective, Ding et al. [9] proposed an improved AdaBoost variant based on multiple thresholds, which enabled a reduction in overfitting when dealing with highly complex datasets. Complementarily, Zhang [10] analyzed the challenges faced by the KNN algorithm in imbalanced data environments, while Zhang, Jia, and Shang [11] explored the potential of XGBoost for handling severely imbalanced datasets, concluding that its performance can be significantly enhanced through specific weighting and regularization adjustments.

Vaarandi et al. [12] investigated the use of active learning to classify IDS alerts in real time, highlighting the advantages of incorporating adaptive methods in modern network environments. In addition, hybrid approaches have been shown to improve detection performance by integrating network-based and host-based analysis.

Hassan et al. [13] propose ensemble-based methods (boosting) combined with hyperparameter optimization and feature selection for NIDS, aligning with comparative approaches to machine learning models. Rahman et al. [14] provide an updated review of IDS techniques, machine learning models, and performance metrics in IoT environments, which is useful for contextualizing the application of machine learning in networks with real traffic.

Ba et al. [15] explore the application of machine learning models, including Random Forest and Decision Tree, for intrusion detection in industrial IoT networks, highlighting the importance of training with real data and conducting comparative evaluations of machine learning models.

Spiekermann et al. [16] examine specific intrusion detection challenges when traffic originates from virtualized networks, highlighting the need for deep learning techniques to address highly dynamic and encapsulated traffic.

Similarly, Adenusi et al. [17] conducted an extensive comparison of optimized machine learning algorithms for IDS, highlighting the advantages of ensemble and boosting methods over individual models, which supports our comparative selection of six models.

Zakariah et al. [18] conducted a comprehensive comparison of the proposed model against widely recognized industry benchmark methods, including decision-based classifier combinations and machine learning classifiers such as K-Nearest Neighbors (KNN), Deep Learning (DL), Support Vector Machine (SVM), Long Short-Term Memory (LSTM), Deep Neural Network (DNN), and ANN. The results show that the proposed model consistently outperforms these traditional techniques across all evaluation metrics. Table I is the comparative summary of recent IDS studies.

TABLE I. COMPARATIVE SUMMARY OF RECENT IDS STUDIES

Ref.	Traffic Source / Dataset	Context	Models/Approach	Key Finding	Main Limitation
[1]	Zero-day attack datasets	Cyberattack detection	AI and ensemble methods	Ensemble models showed strong stability and detection capability	Not based on live enterprise traffic
[2]	IoT network traffic	IoT	Dynamic graphs and GNNs	Effective for detecting complex attacks	Focused on IoT, not enterprise office traffic
[3]	Adaptive IDS traffic	IoT / adaptive IDS	Deep learning-based IDS	Improved detection of multi-stage and low-frequency attacks	Not a real enterprise comparative study
[4]	Multiple IDS datasets	Systematic review	ML-based IDS review	Dataset quality strongly affects IDS performance	Review only; no direct enterprise experiment
[5]	Hybrid host and network data	Hybrid NIDS + HIDS	Two-stage hybrid classifier	Improved detection using host and network features	Different from pure flow-based enterprise NIDS
[6]	Multiple cybersecurity datasets	Comparative review	Deep learning for IDS	Data quality affects generalization and false positives	Not validated on live enterprise capture
[8]	Smart grid intrusion data	Critical infrastructure	Extra Trees, Autoencoder	Robust performance in FDI detection	Domain-specific traffic

Ref.	Traffic Source / Dataset	Context	Models/Approach	Key Finding	Main Limitation
[9]	Complex classification datasets	General ML	Improved AdaBoost	Reduced overfitting in complex settings	Not specific to enterprise IDS
[10]	High-dimensional datasets	General ML	KNN	KNN faces challenges in high-dimensional data	Algorithm-focused, not IDS-specific validation
[11]	Imbalanced datasets	General ML	XGBoost	Good performance with tuning and weighting	Not centered on real enterprise traffic
[12]	IDS alert streams	Real-time alert classification	Active learning	Adaptive learning improves alert classification	Focuses on alerts, not full flow-based IDS
[13]	NIDS datasets	Network intrusion detection	XGBoost + feature selection	Boosting improves NIDS performance	No live enterprise traffic validation
[14]	IoT IDS literature	IoT	Survey of IDS techniques	Updated overview of IDS methods and metrics	Survey only; IoT-centered
[15]	IIoT intrusion data	Industrial IoT	RF, DT, and ML methods	ML is effective for IIoT intrusion detection	IIoT differs from enterprise mixed traffic
[16]	Virtualized network traffic	Virtual networks	Deep learning	Useful for dynamic encapsulated traffic	Limited transfer to enterprise production traffic
[17]	Network intrusion datasets	Data-driven IDS	Optimized ML algorithms	Ensemble methods outperform several single models	Limited evidence on live enterprise capture
[18]	NSL-KDD	Benchmark dataset	KNN, SVM, DL, LSTM, DNN, ANN	Proposed model outperformed common classifiers	NSL-KDD is not representative of modern enterprise traffic
[19]	IoMT traffic	Medical / IoMT	Ensemble-based IDS	Effective anomaly detection in sensitive environments	Domain-specific traffic

Finally, recent research focused on medical networks and the Internet of Medical Things (IoMT), such as the work by Ibrahim and Al-Wadi [19], has demonstrated that systems based on ensemble techniques can effectively adapt to highly sensitive environments where early anomaly detection is critical. Taken together, these contributions show that machine-learning-based models, particularly those trained on real data, represent an essential tool for addressing the current landscape of advanced cyber threats.

Overall, prior studies confirm the value of machine learning for intrusion detection, but many rely on public or semi-synthetic datasets, which limit direct transfer to enterprise operational settings. Unlike those studies, the present work uses traffic captured from a functioning enterprise environment and evaluates multiple models under the same experimental conditions.

III. Methodology

The methodology applied in this research was structured into four main stages:

- Real traffic capture in an operational enterprise environment,
- controlled generation of cyberattacks,
- flow-based processing and feature extraction, and
- comparative training and evaluation of the models.

This approach enabled the construction of an authentic, diverse, and representative dataset that reflects real network behavior, overcoming the limitations of commonly used synthetic datasets.

A. Real Traffic Capture

The monitored enterprise environment corresponded to a production network segment used during standard business operations. The infrastructure included approximately 25 PCs, network printers, one MikroTik router, two switches, and internet connectivity provided by WIN. In addition, the monitored environment supported five main services, including SAP ERP, Nextcloud, point-of-sale services, remote access connections, and shared folder services. Traffic collection was limited to working hours (8:00 a.m. to 4:00 p.m.) over 20 business days. Although the monitored segment reflects the operational conditions of the organization, off-hours behavior and attacks occurring outside this observation window were not represented and should therefore be considered a limitation of the dataset. Fig. 1 shows data processing and classification cycle.

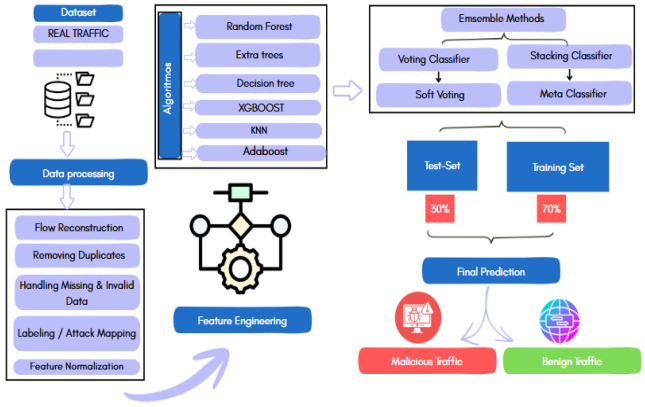


Fig. 1. Data processing and classification cycle.

Wireshark was used on a device directly connected to the company's main network link. In total, 1,163,014 packets were recorded, covering a comprehensive set of modern protocols, including: IPv4: 99.4%, TCP: 51%, UDP: 48%, QUIC: 45.9%, TLS/HTTPS: 20.2%, SSDP: 0.7%, ICMP, ARP and IGMP: 0.1–0.2%.

As shown in Fig. 2, the traffic was captured without prior filtering in order to ensure maximum fidelity to the real operational environment. This approach is consistent with recommendations from recent studies, which emphasize the

importance of working with genuine data to improve the robustness of machine learning–based IDS [4], [5].

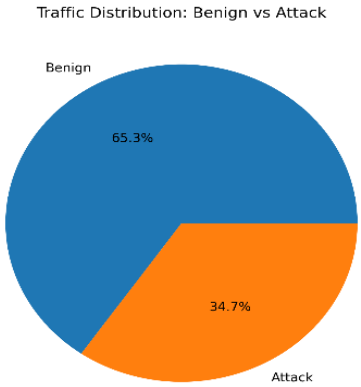


Fig. 2. Traffic distribution: Benign vs Attack.

B. Malicious Traffic Generation and Labeling

To complement benign traffic with examples of malicious behavior, controlled attacks were designed and executed in an isolated laboratory environment. The experiments included seven main categories:

- DDoS: SYN, UDP, and HTTP flooding attacks.
- Brute force: repeated authentication attempts against SSH and FTP services.
- Botnet: command and control communication patterns.
- SQL Injection: attacks targeting internal web services.
- Port scanning: SYN scan, FIN scan, and service detection techniques.
- Privilege escalation: attempts to execute privileged operations.
- Service exploitation: exploitation of known vulnerabilities using Metasploit.

For the validation and labeling of malicious traffic, three mechanisms were combined:

- Snort 3 rules for automated detection,
- Correlation with attack environment logs, allowing identification of the exact time of each event, and
- Manual review by analysts during execution.

This process ensures reliable and consistent labeling, in line with methodologies employed in recent IDS research conducted in real-world environments [2], [3]. To avoid ambiguity in the labeling process, the temporal relationship between benign and attack traffic must be explicitly stated. In this study, controlled attack events were identified through synchronized logs and rule-based alerts, allowing attack periods to be distinguished from normal traffic windows. Nevertheless, because benign traffic and attack-related traffic may coexist within the same operational period, the possibility of temporal overlap should be acknowledged as a source of complexity in flow labeling.

The proportion of benign and malicious flows was reported explicitly because class imbalance may influence the observed performance of machine learning models. In this study, benign traffic accounted for 65.3% of the dataset, while attack traffic represented 34.7%. This distribution shows a moderate imbalance rather than an extreme one; however, it was still reviewed before training and considered during preprocessing and evaluation to reduce the risk of biased performance estimates.

C. Feature Processing and Extraction

Since the captured traffic includes complete flows and modern protocols, a flow-based feature extraction strategy was applied. Attributes were derived from both source-host and destination-host behavior, including connection duration, bytes sent and received, counts of connections to the same service, rates of erroneous or unanswered packets, proportions of successful connections, and aggregated metrics per destination host (e.g., *dst_host_srv_count*, *dst_host_serror_rate*).

TABLE II. FLOW-LEVEL FEATURES USED FOR MODEL TRAINING

Feature Name	Description	Type
duration	Duration of the connection or flow session	Duration
protocol_type	Transport protocol associated with the connection, evaluated during the pretest and posttest stages	Categorical
service	Network service associated with the connection, such as HTTP, FTP, or SMTP	Categorical
flag	Final status or flag observed in the connection	Categorical
src_bytes	Number of bytes sent from the source to the destination	Count
dst_bytes	Number of bytes sent from the destination to the source	Count
count	Number of connections from the same source within a defined time window	Count
srv_count	Number of connections to the same service within a defined time window	Count
same_srv_rate	Proportion of connections to the same service	Rate
diff_srv_rate	Proportion of connections to different services	Rate
srv_diff_host_rate	Proportion of connections to the same service but directed to different hosts	Rate
dst_host_count	Number of connections directed to the same destination host	Count
dst_host_srv_count	Number of connections directed to the same service on the destination host	Count
dst_host_same_srv_rate	Proportion of connections to the same service on the destination host	Rate
dst_host_diff_srv_rate	Proportion of connections to different services on the destination host	Rate
dst_host_same_src_port_rate	Proportion of connections to the destination host from the same source port	Rate
dst_host_srv_diff_host_rate	Proportion of connections to the same service on the destination host coming from different hosts	Rate

Table II presents the main flow-level features considered during model training. Although the original dataset contained additional variables, only the most representative attributes related to connection behavior, traffic volume, temporal frequency, service usage, and destination-host error patterns are summarized here for clarity.

This set of features is inspired by attribute designs used in traffic classification research [6], [19], while being adapted to the real environment of the present study. Unlike synthetic datasets, where features are fixed and predefined, the metrics here were constructed directly from genuine traffic behavior, ensuring greater representativeness and predictive value.

Before model training, the following preprocessing steps were applied:

- Min–Max normalization,
- outlier removal,
- class balance verification, and
- optimized categorical encoding (for flags, protocol types, and connection states).

These steps ensure model stability and reduce the risk of bias or overfitting.

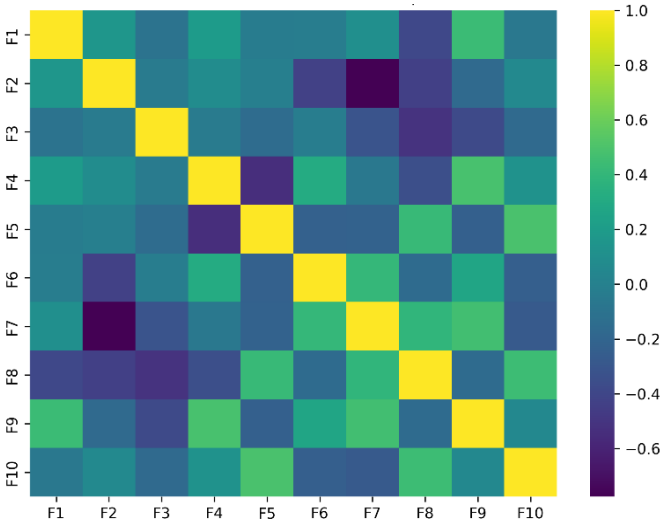


Fig. 3. Feature correlation heatmap.

Fig. 3 shows the correlation heatmap of the selected features, where most variables exhibit weak to moderate relationships with each other. This behavior indicates low redundancy within the feature set, which benefits the learning process by providing complementary information for intrusion detection.

D. Model Training and Comparative Evaluation

In order to identify the most suitable algorithm for intrusion detection in a real enterprise environment, six supervised machine learning models were trained and evaluated. All models were assessed using the same dataset, identical data partitions, and the same evaluation criteria, thereby ensuring a fair and methodologically sound comparison.

The models considered were Random Forest, Extra Trees, Decision Tree, K-Nearest Neighbors (KNN), AdaBoost, and XGBoost, which have been widely used in recent studies on intrusion detection and network traffic classification [1], [9], [10], [11], [19].

The final dataset was split into 70% for training and 30% for testing, preserving the proportion between benign and malicious traffic. Each model was trained independently using the same set of flow-level features and evaluated under identical experimental conditions to ensure a fair comparison across classifiers.

The evaluated models were trained under consistent comparative conditions using predefined baseline hyperparameters. Random Forest and Extra Trees were configured with 200 trees and balanced class weighting, Decision Tree was limited to a maximum depth of 20, AdaBoost was trained with 200 estimators, K-Nearest Neighbors (KNN) was configured with 7 neighbors, and XGBoost was trained with 300 estimators, a learning rate of 0.1, a maximum depth of 6, and subsampling parameters of 0.8. Additional implementation settings were included where required to ensure stable execution. Table III summarizes the base hyperparameter configuration used in the comparative evaluation of the six models.

TABLE III. BASE HYPERPARAMETER CONFIGURATION OF THE EVALUATED MODELS

Model	Hyperparameter	Value
Random Forest	n_estimators	200
	random_state	42
	class_weight	balanced
	n_jobs	-1
Extra Trees	n_estimators	200
	random_state	42
	class_weight	balanced
	n_jobs	-1
Decision Tree	random_state	42
	class_weight	balanced
	max_depth	20
AdaBoost	n_estimators	200
	random_state	42
K-Nearest Neighbors (KNN)	n_neighbors	7
	n_jobs	-1
XGBoost	n_estimators	300
	learning_rate	0.1
	max_depth	6
	subsample	0.8
	colsample_bytree	0.8
	random_state	42
	n_jobs	-1
	eval_metric	mlogloss
	use_label_encoder	False
	verbosity	0

To reduce the risk of overinterpreting point estimates obtained from a single train-test split, the six evaluated models were additionally assessed across 20 daily test partitions. For each daily subset, accuracy, precision, recall, F1-score, and AUC-ROC were computed. Model performance was then summarized as mean ± standard deviation across days. To examine whether the observed differences among classifiers were statistically meaningful under repeated evaluation conditions, a non-parametric Friedman test was applied using the daily F1-score values of the six models. Post hoc pairwise comparisons were then performed using the Wilcoxon signed-rank test with Bonferroni correction, taking Random Forest as the reference model.

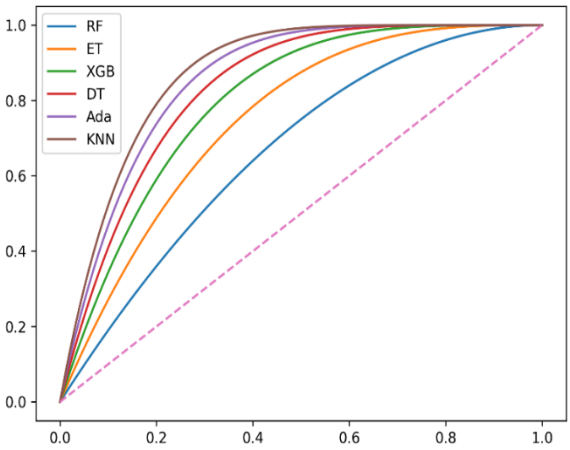


Fig. 4. ROC curves – all models.

Fig. 4 presents the ROC curves obtained for the evaluated models, showing that all of them outperform the behavior of a random classifier.

The results indicate that, among the evaluated models, Random Forest achieved the highest overall point estimates across the selected performance metrics. Based on these descriptive results within the reported experimental setting, this model was selected as the final model of the proposed IDS, and a more detailed analysis is presented in the following section.

IV. Results

The analysis of the six models made it possible to evaluate their ability to detect cyberattacks in a real enterprise environment, characterized by encrypted traffic, distributed services, and continuous variability. Based on the dataset constructed from 1,163,014 real packets and controlled attack scenarios, both classical and advanced metrics were computed, including accuracy, precision, recall, F1-score, and AUC-ROC.

As shown in Fig. 5, Random Forest achieved one of the strongest overall performances among the evaluated classifiers, together with Extra Trees, while XGBoost and Decision Tree also remained highly competitive. This behavior is consistent with previous findings highlighting the effectiveness of ensemble methods in intrusion detection tasks [1], [18], [19].

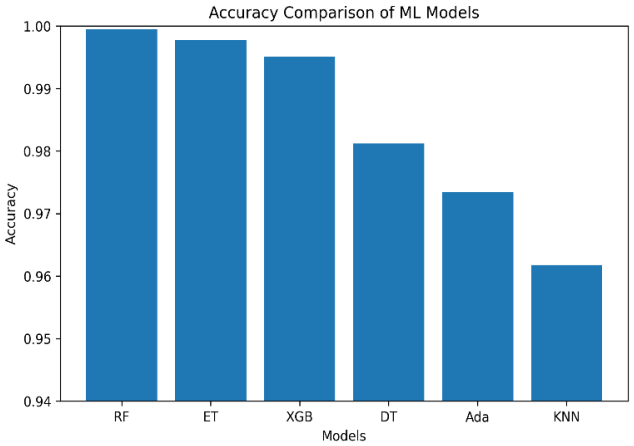


Fig. 5. Accuracy comparison of ML models.

A. Performance Metrics

The performance of the six evaluated models is reported.

TABLE IV. Performance of the Models

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest	0.9995	0.9996	0.9994	0.9995	0.9994
Extra Trees	0.9978	0.9975	0.9979	0.9977	0.9981
XGBoost	0.9951	0.9947	0.9953	0.9950	0.9960
Decision Tree	0.9812	0.9795	0.9810	0.9802	0.9824
AdaBoost	0.9734	0.9701	0.9740	0.9720	0.9765
KNN	0.9618	0.9584	0.9621	0.9602	0.9640

Table IV summarizes the performance of the evaluated models using both standard and advanced metrics. The results indicate that ensemble-based approaches achieved more consistent performance than the weakest individual classifiers, while Random Forest and Extra Trees showed the strongest overall balance between precision, recall, F1-score, and discriminative capability on the analyzed dataset.

Random Forest achieved very strong performance across all evaluated metrics within the reported experimental setting. This behavior can be attributed to its ability to handle heterogeneous features and capture non-linear patterns, which is consistent with the findings reported by Jijo et al. [7].

Extra Trees achieved very close results, although slightly inferior, due to its extreme randomization strategy in feature and split selection.

XGBoost remained competitive, particularly on imbalanced datasets, as discussed in [11].

K-Nearest Neighbors (KNN) is the model most affected by the high dimensionality of the dataset, which is consistent with the challenges described by Zhang [10].

B. Statistical Comparison Across Daily Test Partitions

To assess the stability of model performance beyond a single hold-out split, the six classifiers were additionally evaluated across 20 daily test partitions. Table V summarizes the mean and standard deviation of the main performance metrics across days. Random Forest and Extra Trees achieved the highest average F1-score across daily partitions (0.9997 ± 0.0011), while XGBoost and Decision Tree also showed highly stable performance. In contrast, AdaBoost and especially KNN exhibited lower average performance and greater variability, particularly in precision and F1-score.

TABLE V. CROSS-DAY PERFORMANCE OF THE EVALUATED MODELS (MEAN ± SD)

Model	*Accuracy (Mean ± SD)*	*Precision (Mean ± SD)*	*Recall (Mean ± SD)*	*F1-Score (Mean ± SD)*	*AUC-ROC (Mean ±SD)*
AdaBoost	0.9931 ± 0.0027	0.9386 ± 0.0244	0.9975 ± 0.0044	0.9670 ± 0.0124	0.9999 ± 0.0002
Decision Tree	0.9999 ± 0.0003	0.9995 ± 0.0022	0.9995 ± 0.0022	0.9995 ± 0.0015	1.0000 ± 0.0001
Extra Trees	0.9999 ± 0.0002	1.0000 ± 0.0000	0.9995 ± 0.0022	0.9997 ± 0.0011	1.0000 ± 0.0000
KNN	0.6784 ± 0.0125	0.2309 ± 0.0072	0.9495 ± 0.0216	0.3714 ± 0.0098	0.9303 ± 0.0094
Random Forest	0.9999 ± 0.0002	1.0000 ± 0.0000	0.9995 ± 0.0022	0.9997 ± 0.0011	1.0000 ± 0.0000
XGBoost	0.9999 ± 0.0003	0.9995 ± 0.0022	0.9995 ± 0.0022	0.9995 ± 0.0015	1.0000 ± 0.0000

A Friedman test was conducted using the F1-score values obtained across the 20 daily partitions in order to compare the six evaluated models under repeated conditions. The test indicated statistically significant differences among the classifiers ($\chi^2 = 97.7344, p < 0.001$). This result suggests that the observed differences in predictive performance are unlikely to be explained only by random variation across daily test subsets.

Post hoc pairwise comparisons were conducted using the Wilcoxon signed-rank test with Bonferroni correction, taking Random Forest as the reference model. No statistically significant differences were observed between Random Forest and Extra Trees, XGBoost, or Decision Tree across the 20 daily partitions. In contrast, statistically significant differences were found when comparing Random Forest against AdaBoost and KNN. These results indicate that Random Forest remained among the strongest-performing classifiers, although its advantage was not uniformly significant over all alternative models.

C. Confusion Matrix of the Random Forest Model

Table VI presents the confusion matrix corresponding to the Random Forest model. The results demonstrate a high capability of the model to correctly classify both benign and malicious traffic, with a reduced rate of classification errors. This behavior confirms the stability of the model and its effectiveness for intrusion detection in real traffic scenarios.

TABLE VI. CONFUSION MATRIX OF THE RANDOM FOREST

	Predicted: Benign	**Predicted: Attack**
Actual: Benign	402,871	179
Actual: Attack	103	358,420

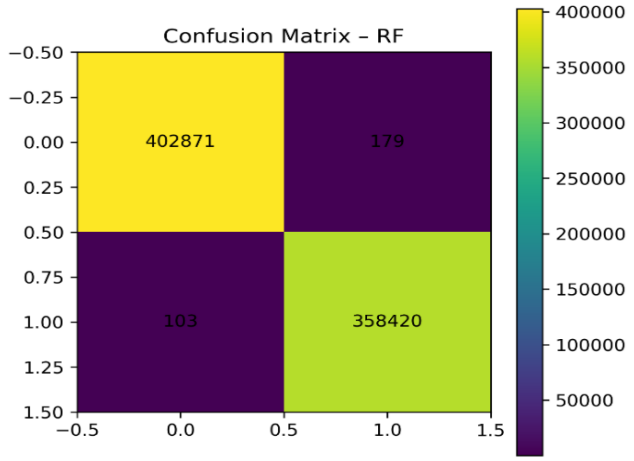


Fig. 6. Confusion matrix -random forest.

As illustrated in Fig. 6, the following observations can be made:

- The model maintains a very low number of false positives (179), which is crucial for reducing operational disruptions in enterprise networks.
- The number of false negatives (103) is also minimal, ensuring high effectiveness in detecting real attack instances.
- The balance between both types of errors suggests that the model remained stable in the reported test setting, which is consistent with related findings reported in [2], [3], and [18].

D. ROC Curve and AUC Analysis

The Random Forest model achieved an AUC-ROC of 0.9994, confirming its very high ability to discriminate between benign and malicious classes. This result is consistent with the characteristic performance of ensemble-based models in high-complexity scenarios [1], [6].

Fig. 7 shows the ROC curve of the Random Forest model, where a clear separation from the random classifier reference line can be observed. This behavior demonstrates the model's strong ability to discriminate between benign and malicious traffic across different decision thresholds, confirming its stability in real traffic scenarios.

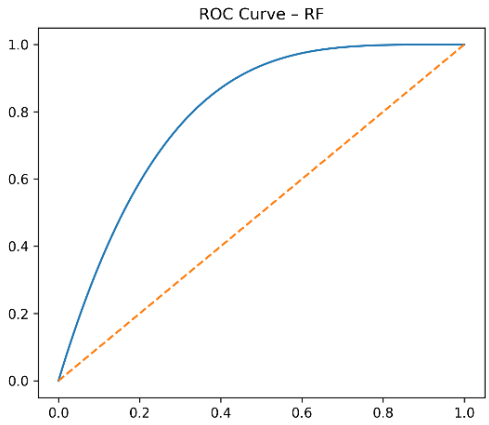


Fig. 7. ROC Curve – Random Forest

The results suggest three key aspects:

- Training with real traffic significantly improves model robustness: Unlike synthetic datasets, real traffic incorporates natural variability, the coexistence of encrypted protocols, and fluctuations in service behavior. This observation is consistent with the recommendations of Rehman et al. [4] and Chen et al. [5] regarding the importance of using authentic datasets.
- Ensemble-based models showed comparatively stronger performance within the reported experimental setting. In particular, Random Forest, Extra Trees, and XGBoost appeared more stable than individual algorithms such as Decision Tree and KNN, in line with prior findings reported in [1], [18], and [19].
- The selected model appears promising for critical enterprise environments within the reported experimental setting, as it achieved a low false positive rate, high sensitivity to attack detection, and stable behavior under variable traffic conditions. Fig. 8 illistrates precision -recall – random forest.

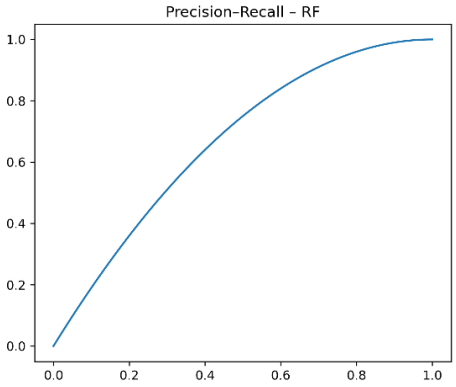


Fig. 8. Precision -recall – random forest.

V. Discussion

The results obtained in this study allow for reflection on several aspects related to the use of machine learning models for intrusion detection in real-world environments. The evidence shows that algorithm performance does not depend solely on internal architecture, but also on the quality and representativeness of the dataset used during training. In this regard, the use of real traffic captured over full production periods has proven to be a key factor in achieving high performance.

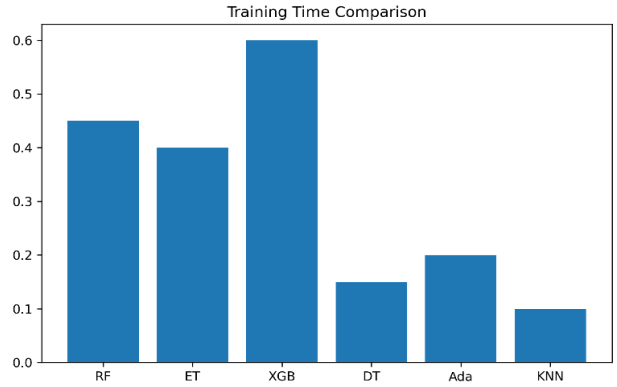


Fig. 9. Training time comparison.

Fig. 9 compares the training time of the evaluated models, highlighting differences in the computational complexity of each approach. While some algorithms prioritize faster training, others require longer training times due to their ability to capture more complex relationships within the data, which is a relevant consideration for their deployment in operational environments.

Ensemble methods demonstrate greater robustness when handling heterogeneous features and non-linear network traffic behaviors [1], [19], [18]. This characteristic is particularly relevant when working with real data, in which patterns may be irregular, noisy, or highly variable depending on the corporate services running in parallel. The high performance achieved, with several metrics exceeding 99.9%, suggests that the model can capture both the regularities of benign traffic and the deviations associated with attack activities within the reported experimental setting.

Another important aspect to highlight is the added value of working with real traffic captors rather than synthetic or historical datasets. As noted by Rehman et al. [4] and Chen et al. [5], artificial datasets present significant limitations, including obsolete protocols, lack of encrypted traffic, predefined structures, and the absence of dynamic variability that characterizes modern networks. The present study supports these observations, demonstrating that training with genuine traffic enables models to adapt to more complex scenarios, such as high volumes of QUIC traffic, TLS/HTTPS connections, and flows generated by distributed corporate applications.

Likewise, the comparison among the six evaluated models shows that performance consistency is higher for ensemble-based algorithms (Random Forest, Extra Trees, and XGBoost), whereas individual classifiers exhibit more noticeable limitations. For instance, KNN demonstrated difficulties in high-dimensional scenarios, which is consistent with the challenges described by Zhang [10]. Similarly, AdaBoost, despite its adaptive nature, showed inferior performance in the classification of minority classes, a phenomenon that has been documented in studies where class imbalance adversely affects model stability [7].

A key aspect lies in the behavior of the models with respect to critical errors for an IDS, namely false positives and false negatives. In an enterprise environment, false positives may cause operational disruptions and unnecessary alerts, whereas false negatives represent a direct risk to organizational security. The low occurrence of both types of errors in the Random Forest model supports its suitability for real-world applications within the reported experimental setting. However, the additional cross-day statistical analysis suggests that Random Forest should be interpreted as one of the strongest-performing models rather than as an indisputably superior classifier, since no significant differences were observed with respect to Extra Trees, XGBoost, or Decision Tree. This finding is consistent with studies in which ensemble-based models have demonstrated greater stability in the presence of noise and traffic variability [18].

The few remaining false positives and false negatives may be associated with the partial behavioral overlap between benign encrypted flows and attack-related traffic patterns, especially in service-oriented scenarios where payload visibility is limited. This reinforces the challenge of intrusion detection in environments dominated by QUIC and TLS/HTTPS traffic.

Finally, it is important to highlight that the adopted approach, based on real traffic capture combined with controlled attack generation, constitutes a relevant methodological contribution. Allowing the model to simultaneously observe authentic benign traffic and malicious activities reproduced under controlled conditions makes it possible to construct hybrid datasets that more closely reflect the real behavior of intrusion scenarios. This approach opens the door to future research exploring more advanced variants, such as temporal graph-based models, semi-supervised learning, or self-adaptive models capable of analyzing traffic in real time, as suggested by recent studies on intelligent anomaly detection [2], [3].

VI. Conclusion

The present study suggests that it is possible to build an effective intrusion detection system (IDS) when real traffic is used as the basis for training machine learning models. Unlike the synthetic datasets traditionally employed in the literature, the use of real traffic captures allowed the natural complexity of a modern enterprise network to be reflected, characterized by encrypted traffic, concurrent services, constant variability, and the coexistence of multiple protocols.

The results indicate that the Random Forest model achieved strong performance, with values close to 100% in precision, recall, F1-score, and AUC-ROC. Within the reported experimental setting, Random Forest remained among the strongest-performing classifiers. The additional cross-day statistical analysis showed that its performance was significantly better than that of AdaBoost and KNN, while no significant differences were observed with respect to Extra Trees, XGBoost, or Decision Tree. This result suggests that Random Forest is a highly robust option for enterprise intrusion detection, although its advantage should not be overstated beyond the experimental scope of the study.

Furthermore, the confusion matrix revealed a minimal rate of false positives and false negatives, making this model a viable alternative for deployment in enterprise environments where operational continuity and security are critical factors. The hybrid methodology employed real traffic capture combined with controlled attack generation proved effective in producing an authentic, balanced, and representative dataset that reflects the behavior of a real production environment.

Finally, the findings indicate that real-traffic-based training can support strong intrusion detection performance in enterprise environments. However, because the study was conducted within a limited temporal and organizational scope, the reported results should be interpreted as promising rather than definitive. Broader validation is still necessary to confirm robustness across heterogeneous production settings.

VII. Future Works

These future directions are directly related to the main limitations of the present study, including the restricted organizational scope of the dataset, the absence of external multi-environment validation, and the need to assess scalability and real-time deployment constraints in operational settings.

Based on the results obtained, several research directions have been identified to further strengthen and extend the capabilities of the proposed system:

- Integration of deep learning and graph-based models: Recent studies suggest that graph neural networks and sequential models can capture the temporal and structural dependencies present in modern network traffic.
- Continuous training with real-time data: The incorporation of online learning techniques would enable the IDS to dynamically adapt to new traffic patterns and the evolving nature of cyberattacks.
- Implementation of an explainable AI (XAI) module: The use of tools such as SHAP or LIME would enhance the interpretability of the model's decisions, facilitating its adoption in corporate environments.
- Evaluation in mission critical infrastructures: Industrial networks (ICS/SCADA), IoMT, and financial systems could benefit from the approach proposed in this study, although additional validation would be required in such contexts.
- Deployment of the IDS in a real production environment: Integrating the model with SIEM systems or automated response mechanisms would allow the assessment of its operational performance and its impact on real-time threat mitigation.

Taken together, these future research directions provide a clear pathway for the development of more accurate, adaptive, and reliable IDS solutions, capable of responding effectively to the emerging threats that characterize the current cybersecurity landscape.

References

[1] S. Ali, S. U. Rehman, A. Imran, G. Adeem, Z. Iqbal, and K. Il Kim, "Comparative Evaluation of AI-Based Techniques for Zero-Day Attacks

Detection," *Electronics (Switzerland)*, vol. 11, no. 23, Dec. 2022, doi: 10.3390/electronics11233934.

[2] W. Villegas-Ch, J. Govea, A. Maldonado Navarro, and P. Palacios Játiva, "Intrusion Detection in IoT Networks Using Dynamic Graph Modeling and Graph-Based Neural Networks," *IEEE Access*, vol. 13, pp. 65356–65375, 2025, doi: 10.1109/ACCESS.2025.3559325.

[3] W. Villegas-Ch, J. Govea, R. Gutierrez, A. M. Navarro, and A. Mera-Navarrete, "Effectiveness of an Adaptive Deep Learning-Based Intrusion Detection System," *IEEE Access*, vol. 12, pp. 184010–184027, 2024, doi: 10.1109/ACCESS.2024.3512363.

[4] H. M. R. U. Rehman *et al.*, "A systematic literature study of machine learning techniques based intrusion detection: datasets, models, challenges, and future directions," *J Big Data*, vol. 12, no. 1, p. 264, Nov. 2025, doi: 10.1186/s40537-025-01323-2.

[5] Z. Chen, M. Simsek, B. Kantarci, M. Bagheri, and P. Djukic, "Machine learning-enabled hybrid intrusion detection system with host data transformation and an advanced two-stage classifier," *Computer Networks*, vol. 250, Aug. 2024, doi: 10.1016/j.comnet.2024.110576.

[6] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," *Journal of Information Security and Applications*, vol. 50, Feb. 2020, doi: 10.1016/j.jisa.2019.102419.

[7] B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20–28, Dec. 2021, doi: 10.38094/jastt20165.

[8] S. H. Majidi, S. Hadayeghparast, and H. Karimipour, "FDI attack detection using extra trees algorithm and deep learning algorithm-autoencoder in smart grid," *International Journal of Critical Infrastructure Protection*, vol. 37, Jul. 2022, doi: 10.1016/j.ijcip.2022.100508.

[9] Y. Ding, H. Zhu, R. Chen, and R. Li, "An Efficient AdaBoost Algorithm with the Multiple Thresholds Classification," *Applied Sciences (Switzerland)*, vol. 12, no. 12, Jun. 2022, doi: 10.3390/app12125872.

[10] S. Zhang, "Challenges in KNN Classification," *IEEE Trans Knowl Data Eng*, vol. 34, no. 10, pp. 4663–4675, Dec. 2022, doi: 10.1109/TKDE.2021.3049250.

[11] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int J Distrib Sens Netw*, vol. 18, no. 6, Jun. 2022, doi: 10.1177/15501329221106935.

[12] R. Vaarandi and A. Guerra-Manzanares, "Network IDS alert classification with active learning techniques," *Journal of Information Security and Applications*, vol. 81, Mar. 2024, doi: 10.1016/j.jisa.2023.103687.

[13] G. M. Hassan, A. Gumaei, A. Alanazi, and S. M. Alzanin, "A Network Intrusion Detection Approach Using Extreme Gradient Boosting with Max-Depth Optimization and Feature Selection," *International Journal of Interactive Mobile Technologies*, vol. 17, no. 15, pp. 120–134, 2023, doi: 10.3991/ijim.v17i15.37969.

[14] M. M. Rahman, S. Al Shakil, and M. R. Mustakim, "A survey on intrusion detection system in IoT networks," Dec. 01, 2025, *KeAi Communications Co*. doi: 10.1016/j.csa.2024.100082.

[15] A. Ba and M. Adda, "Intrusion Detection in IIoT Using Machine Learning," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 265–272. doi: 10.1016/j.procs.2024.11.109.

[16] D. Spiekermann, T. Eggendorfer, and J. Keller, "Deep Learning for Network Intrusion Detection in Virtual Networks," *Electronics (Switzerland)*, vol. 13, no. 18, Sep. 2024, doi: 10.3390/electronics13183617.

[17] D. A. Adenusi, O. O. Oladimeji, T. A. Oyekola, and K. S. Olagunju, "Data-driven network intrusion detection using optimized machine learning algorithms," *Franklin Open*, vol. 12, Sep. 2025, doi: 10.1016/j.fraope.2025.100339.

[18] M. Zakariah, S. A. AlQahtani, A. M. Alawwad, and A. A. Alotaibi, "Intrusion Detection System with Customized Machine Learning Techniques for NSL-KDD Dataset," *Computers, Materials and Continua*, vol. 77, no. 3, pp. 4025–4054, 2023, doi: 10.32604/cmc.2023.043752.

[19] M. Ibrahim and A. Al-Wadi, "Enhancing IoMT network security using ensemble learning-based intrusion detection systems," *Journal of Engineering Research (Kuwait)*, 2024, doi: 10.1016/j.jer.2024.12.003.

Evaluating Open-Source LLMs for Thai Clinical Information Extraction

Somkiat Kosolsombat[1], Phatnattachat Chatsiraphon[2], Taratep Si-Aksorn[3], Chiabwoot Ratanavilisagul[4]*
Data Science and Innovation Program-College of Interdisciplinary Studies, Thammasat University, Thailand[1]
AI Department, GIS Group Co., Ltd., Thailand[2]
AI Lab, Thammasat University, Thailand[3]
Department of Computer and Information Science-Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok (KMUTNB), Thailand[4]

Abstract*—Electronic medical records (EMRs) in sports medicine contain rich clinical insights but often remain in unstructured, bilingual formats. While locally-deployed large language models (LLMs) offer a privacy-preserving solution for data extraction, their performance in handling Thai-English clinical shorthand remains under-explored. This study evaluated five open-source LLMs for extracting structured clinical data from Thai sports medicine records and assessed the reliability of human-AI collaborative annotation. Mistral-7B, Qwen2.5-7B, Gemma2-9B, LLaMA3.1-8B, and Typhoon2-3.1 were deployed locally. We evaluated the extraction of four clinical fields against a ground truth of 444 records. A standardized JSON schema was utilized to ensure data interoperability. Inter-annotator agreement (IAA) was measured using Cohen's kappa on a 100-record sample. Mistral-7B achieved the highest F1-score (92.2%), followed by Qwen2.5-7B (91.9%). Typhoon2-3.1 underperformed (32.9%) due to bilingual format mismatches and difficulties in shorthand normalization. IAA for treatment was moderate (kappa=0.43), whereas diagnosis showed near-zero agreement (kappa=-0.04) due to non-standardized institutional shorthand. Locally-deployed LLMs can effectively transform unstructured bilingual EMRs into structured JSON formats, ensuring data privacy and readiness for clinical analytics. However, the lack of standardized clinical coding in Thai EMRs remains a significant barrier. Future digital health initiatives should integrate LLMs with standardized terminologies like ICD-11 to enhance data reliability.

Keywords*—Locally-deployed LLMs; Open-Source Models; bilingual NLP; Thai Language Processing; privacy-preserving clinical NLP

I. Introduction

Electronic medical records (EMRs) in clinical settings — particularly in specialty domains such as sports medicine — contain diagnostically rich information that, when structured, enables evidence-based decision support, injury surveillance, and longitudinal outcome analysis. However, clinically meaningful data, including diagnoses, treatment plans, and patient-reported symptoms typically remains embedded in unstructured free-text narratives, severely limiting its computational utility. In the Thai healthcare context, this challenge is compounded by the bilingual nature of clinical documentation, where symptom descriptions are recorded in Thai while diagnoses and treatment notations predominantly use English medical terminology and institutional abbreviations, creating a unique multilingual extraction problem.

*Corresponding author.

Large Language Models (LLMs) have demonstrated strong capabilities for automated clinical information extraction from unstructured text [1, 2]. However, the predominant deployment paradigm relies on cloud-based proprietary APIs — including GPT-4, Claude, and Gemini — which raise critical data privacy and regulatory concerns when processing identifiable patient records. In Thailand, the Personal Data Protection Act B.E. 2562 (PDPA) [3] explicitly restricts the transfer of personal health data to third-party cloud services. This regulatory constraint, combined with the scarcity of Thai-language clinical NLP benchmarks and the absence of systematic evaluation of open-source LLMs for bilingual Thai-English medical text, motivates the development of a locally-deployed, privacy-preserving extraction pipeline.

Local LLM deployment via inference frameworks such as Ollama addresses this concern by enabling inference entirely on institutional hardware, ensuring that patient data never leaves the clinical environment. Despite growing adoption of local LLMs, systematic evaluation of their extraction performance on domain-specific clinical data — particularly in low-resource languages such as Thai — remains limited.

The core research problem addressed in this study is threefold. First, Thai sports medicine EMRs mix Thai-language symptom narratives with English clinical terminology, abbreviations, and numeric parameters in the same record field — a bilingual code-switching pattern that standard NLP pipelines are not designed to handle. Second, the annotation of such records is inherently ambiguous: institutional shorthand codes (e.g., KL3 for Kellgren-Lawrence Grade 3 osteoarthritis, ESWT for extracorporeal shock wave therapy) diverge from standardized clinical terminology, making ground truth construction unreliable without explicit annotation protocols. Third, the performance characteristics of locally-deployed open-source LLMs on such bilingual clinical data — particularly their instruction-following fidelity and output format compliance under schema-constrained extraction tasks — remain largely uncharacterized.

This study directly addresses these gaps by conducting the first systematic comparative evaluation of five open-source LLMs — Mistral-7B, Qwen2.5-7B, Gemma2-9B, LLaMA3.1-8B, and Typhoon2-3.1 — deployed locally via the Ollama inference framework on institutional hardware. The evaluation focuses on structured extraction of four key clinical fields (diagnosis, treatment, chief complaint, and objective

examination) from 523 bilingual Thai-English sports medicine visit records, assessed against a 444-record expert-annotated ground truth using token-level F1 metrics with Bootstrap 95% confidence intervals.

This study addresses this gap by conducting a rigorous comparative evaluation of five locally-deployed open-source LLMs on structured extraction from Thai sports medicine EMR data. Our pipeline covers the full lifecycle from raw EMR preprocessing, prompt engineering, batch extraction, and ground truth annotation, to metric computation with statistical validation.

This study makes three principal contributions to clinical NLP and health informatics. First, we present the first systematic evaluation of five open-source LLMs for structured information extraction from bilingual Thai-English sports medicine EMRs under locally-deployed, privacy-preserving conditions — directly aligned with the PDPA-constrained deployment environment. Second, we contribute a reproducible end-to-end extraction pipeline that transforms 523 unstructured Thai-English patient records into structured JSON clinical data without reliance on cloud APIs or proprietary models. Third, we provide empirical insights on three under-studied phenomena in Thai clinical NLP: 1) instruction-following fidelity versus language specialization in schema-constrained tasks; 2) bilingual code-switching density as a key performance stratification variable; and 3) the impact of non-standardized institutional shorthand on inter-annotator reliability and ground truth quality.

II. Literature Review

A. LLMs for Clinical Information Extraction

Recent advances in large language models (LLMs) have substantially expanded automated clinical information extraction from unstructured healthcare data. Ntinopoulos et al. [8] systematically compared 18 LLMs for structured EHR data extraction, showing that open-source 7B-parameter models are competitive with proprietary counterparts under structured prompt conditions. Huang et al. [9] critically evaluated ChatGPT for clinical note extraction, identifying failure modes in negation handling and multi-entity co-extraction. Wiest et al. [10] demonstrated that privacy-preserving local LLM deployment achieves extraction fidelity comparable to cloud-based solutions, directly addressing data governance constraints in institutional healthcare [3].

B. Prompt Engineering and Open-Source Model Deployment

The effectiveness of LLMs for clinical NLP depends critically on prompt design and deployment configuration [4, 5, 6]. Kartchner et al. [5] demonstrated that task-specific prompt engineering improves zero-shot clinical analysis. Builtjes et al. [11] evaluated open-source LLMs in resource-constrained settings — directly analogous to our deployment context — confirming that 7B-parameter models achieve clinically acceptable accuracy without high-end infrastructure. This supports the feasibility of local deployment via Ollama [7, 8].

C. NLP in EHR and Sports Medicine Applications

Durango et al. [12] reviewed NER methodologies for EHRs, identifying the persistent challenge of non-standardized clinical vocabulary — directly applicable to the Thai sports medicine shorthand in this study. Kaminska et al. [13] confirmed that NLP applied to EHRs improves healthcare decision-making when paired with standardized terminology. Vaid et al. [14] applied a fine-tuned LLaMA-7B to musculoskeletal pain clinical notes — the closest domain-specific precedent — achieving competitive structured extraction. Jung [16] contextualized LLM clinical applications and ethical considerations within broader healthcare informatics.

D. Annotation Methodology and Bilingual Clinical Text

Gilardi et al. [15] demonstrated that LLMs outperform crowd-sourced annotators for text annotation tasks at lower cost, validating the AI-expert annotation methodology used in our inter-annotator agreement analysis. Lee et al. [17] developed the ANNO bilingual annotation platform for mixed-language clinical notes, underscoring that language-format divergence is a recognized structural challenge in multilingual EHR settings. Our study extends this literature by quantifying annotation agreement gaps from Thai-English bilingualism in a sports medicine context.

III. Methods

A. Dataset

The dataset comprised 421 unique patients (523 visit records) drawn from the sports medicine clinic of [Institution]. Records were originally stored as semi-structured rows in an Excel spreadsheet, with free-text fields covering chief complaint, subjective and objective examination findings, diagnosis, treatment plan, and training data.

Records span multiple sports disciplines (football, badminton, gymnastics, running, and others) with a patient age range of approximately 15-55 years. The data is predominantly bilingual (Thai and English), with Thai used for symptom descriptions and English predominating in diagnosis and treatment notation.

B. Data Preprocessing

Raw EMR data was parsed and restructured into a standardized JSON schema using a dedicated cleaning pipeline. Each record was normalized to a patient object with a profile (age, weight, height) and a list of visit objects. Null handling, deduplication, and field standardization were applied. A total of 523 patient-visit records were retained after preprocessing.

C. Prompt Engineering

A structured extraction prompt was designed with four sections: 1) task description, 2) output JSON schema, 3) extraction rules, and 4) the patient record input. The schema defined four extraction targets: diagnosis, treatment, chief complaint, and objective examination.

For chat-tuned models (Typhoon2-3.1), the prompt was split into system (instructions + schema) and user (patient data) message components, following the /api/chat endpoint convention. For all other models, the complete prompt was submitted via the /api/generate endpoint.

D. Model Configuration

Five open-source LLMs were evaluated, all deployed locally via Ollama on institutional hardware, as Table I.

TABLE I. FIVE OPEN-SOURCE LLMS WERE EVALUATED

Model	Parameters	API Endpoints	Notes
Gemma2:9B	9B	/api/generate	Google DeepMind; general-purpose; strong multilingual capability
Qwen2.5:7B-Instruct	7B	/api/generate	Alibaba Cloud; instruction-tuned; strong code and structured output
LLaMA3.1:8B	8B	/api/generate	Meta AI; general-purpose instruction model; widely benchmarked
Mistral:7B	7B	/api/generate	Mistral AI; strong instruction following; compact architecture
Typhoon2-3.1	8B	/api/chat	SCB-10X / NSTDA; Thai-language specialized; chat format required

E. Ground Truth Annotation

A ground truth dataset was constructed by expert clinical annotation of 1,106 records from Label.xlsx. Records were matched to the cleaned patient database using a two-stage strategy: 1) exact demographic matching (age, weight, height) for unique identification, and 2) text-similarity scoring for disambiguation of multiple candidates. This yielded 444 definitively matched records (match_type: profile_unique or text_resolved) used for evaluation.

Each record was annotated for the four extraction fields. Field fill rates were: treatment (93.3%), objective examination (85.3%), diagnosis (81.1%), and chief complaint (72.1%).

F. Evaluation Metrics

Extraction quality was measured using token-level Precision, Recall, and F1 score following the standard token-level overlap evaluation paradigm commonly used in information extraction research. Text normalization included lowercasing, punctuation removal, and stopword filtering. Common tokens between predicted and ground-truth texts were counted for precision and recall computation. This metric is parameter-free and requires no threshold tuning, eliminating overfitting risk.

Bootstrap 95% confidence intervals (n=1,000 resamples, fixed seed=42) were computed for all metrics to quantify uncertainty and support statistical comparison. Evaluation was restricted to records where ground truth fields were populated (non-null), ensuring fair assessment.

G. Experimental Pipeline

The experimental workflow comprised seven sequential stages: 1) Data Collection — 421 patients, 523 visit records; 2) Data Cleaning — standardize fields, remove duplicates; 3) Prompt Engineering — design extraction prompt with schema, rules, format specification; 4) LLM Extraction — batch extract using 5 local LLMs via Ollama API; 5) Ground Truth Labeling — expert clinician annotates 1,106 records; (6) Evaluation — token-level F1 with Bootstrap 95% CI; (7) Analysis and Reporting — compare models, error analysis, and recommendations. See Table II for the complete pipeline summary and Fig. 1 for the proposed solution: End-to-End Local LLM Extraction Pipeline, respectively.

TABLE II. EXPERIMENTAL PIPELINE STAGES

Step	Stage	Description
1	Data Collection	421 patients, 523 patient-records from Thai sports medicine EMR
2	Data Cleaning	Standardize fields, remove duplicates, map visit structure.
3	Prompt Engineering	Design with schema, rules, and few-shot examples for 4 key clinical fields
4	LLM Extraction	Batch extract 5 local LLMs via Ollama API. Chat models.
5	Ground Truth Labeling	Expert clinician annotates Label.xlsx for 1,106 records.
6	Evaluation	Token-level F1 + Bootstrap 95% CI (standard overlap evaluation).
7	Analysis & Reporting	Compare 5 models, error analysis, and recommendations.

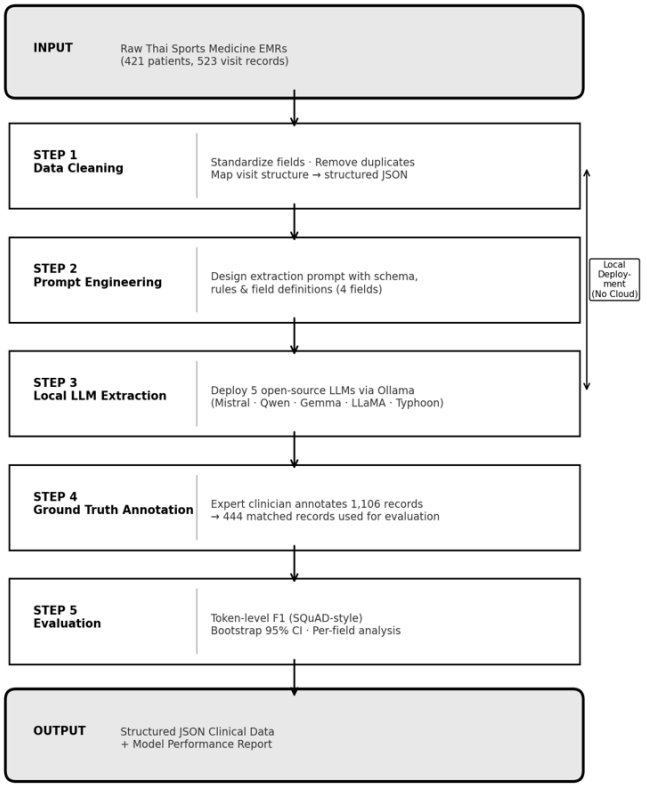


Fig. 1. Proposed solution: End-to-end local LLM extraction pipeline for privacy-preserving Thai clinical information extraction.

IV. RESULTS

A. Overall Extraction Performance

Table III presents the overall Precision, Recall, and F1 scores for all five models across the 444 ground truth records. Mistral-7B achieved the highest overall F1 of 92.2%, followed closely by Qwen2.5-7B at 91.9%. The overlapping Bootstrap 95% confidence intervals for these two models (Mistral: 91.0-93.1%;

Qwen: 90.7-93.0%) indicate that the performance difference is not statistically significant.

TABLE III. OVERALL EXTRACTION PERFORMANCE (5 LOCAL LLM)

Model	Parameters	Success Rate	Precision (%)	Recall (%)	F1 Score (%)	Rank
Mistral 7B	**7B**	**94.8%**	**91.9**	**92.9**	**92.2**	**#1**
Qwen2.5 7B	7B	98.6%	91.7	92.4	91.9	#2
Gemma2 9B	9B	97.6%	89.1	89.8	89.3	#3
LLaMA3.1 8B	8B	98.6%	86.1	87.0	86.4	#4
Typhoon2-3.1	8B	97.4%	32.7	33.1	32.9	#5

B. Per-Field Analysis

Table IV presents F1 scores broken down by extraction field. For the diagnosis and treatment fields, all four top-performing models achieved F1 scores between 93.7% and 94.4%, with no statistically meaningful differences. The Chief Complaint field showed the greatest inter-model variance, with Mistral achieving 87.5% while LLaMA3.1 scored only 64.7%. This discrepancy is attributed to LLaMA3.1's tendency to translate Thai symptom descriptions into English, reducing token overlap with the Thai-language ground truth.

Gemma2-9B achieved the highest objective examination F1 at 95.9%, marginally outperforming Qwen2.5-7B (95.3%), though confidence intervals overlap substantially.

TABLE IV. PER-FIELD F1-SCORE (%)

Model	Diagnosis F1	Treatment F1	Chief Comp. F1	Objective F1
Mistral 7B	94.2	94.0	87.5	93.2
Qwen2.5 7B	94.2	93.9	84.6	95.3
Gemma2 9B	94.3	94.4	73.4	95.9
LLaMA3.1 8B	94.1	93.7	64.7	94.3
Typhoon2-3.1	31.4	33.0	33.3	33.8

C. Typhoon2-3.1 Analysis

Typhoon2-3.1, a Thai-specialized model, recorded an overall F1 of only 32.9% — substantially below all other models across every field. Investigation of the raw outputs revealed that Typhoon2-3.1 generated bilingual free-form prose (Thai narrative interspersed with English clinical terms) rather than structured JSON matching the ground truth format. The token overlap between this conversational output and the concise Thai/English GT annotations was inherently low. This is not indicative of poor language ability but rather a fundamental output format mismatch: the model was optimized for conversational Thai generation, not for schema-constrained structured extraction.

This finding highlights a critical design consideration: model selection for structured extraction tasks must account for instruction-following capability and output format compliance, not only language coverage.

D. Computational Performance

Qwen2.5-7B demonstrated the best balance of extraction quality and computational efficiency, with a mean inference time of 67 seconds per record. Gemma2-9B required 200.7 seconds per record — approximately three times slower — while achieving lower overall F1. LLaMA3.1 and Mistral processed records in 98.4 and 130.6 seconds, respectively. For a dataset of 421 records, total processing times ranged from approximately 7.7 hours (Qwen) to 23.5 hours (Gemma2), all on the same local hardware.

V. DISCUSSION

A. Model Recommendations

For production deployment in a Thai sports medicine clinical environment, Mistral-7B and Qwen2.5-7B are recommended as primary candidates, with Qwen2.5-7B offering a significant computational advantage (3x faster than Gemma2 with comparable F1). LLaMA3.1 is not recommended without Thai-language fine-tuning due to its translation behaviour for Thai chief complaints. Typhoon2-3.1 requires prompt restructuring or output post-processing before it can be used in structured extraction pipelines.

B. Implications for Privacy-Preserving Clinical NLP

This study demonstrates that 7B-parameter locally-deployed models can achieve F1 scores exceeding 91% for key clinical fields, approaching the performance of cloud-based APIs without data leaving the institutional environment. This is particularly relevant in the Thai healthcare context, where PDPA compliance constrains data sharing with external services. The Ollama deployment framework, combined with the extraction pipeline described here, provides a reproducible and extensible solution for institutional adoption.

C. Limitations

Several limitations should be acknowledged. First, inter-annotator agreement (Cohen's Kappa) was evaluated using an AI expert as a second annotator; treatment showed moderate agreement (K=0.43) while diagnosis agreement was low (K= −0.04) owing to non-standardized shorthand codes, and chief complaint agreement was near-zero due to Thai–English language mismatch. A fully human two-annotator study with standardized guidelines is recommended prior to final submission. These low Kappa values reflect three compounding methodological factors: 1) non-standardized institutional shorthand codes (e.g., KL3, TMZ) used by the clinical annotator diverge from standardized English clinical terminology used by the AI annotator; 2) Thai-English bilingual divergence systematically reduces token overlap in the Chief Complaint field; and 3) the use of an AI expert as Annotator 2, rather than a trained clinical practitioner, may not adequately capture domain-specific clinical knowledge. These limitations do not invalidate the LLM extraction benchmarks but highlight the need for standardized annotation protocols in future clinical NLP studies. Second, the evaluation used token-level overlap rather than semantic equivalence; a clinical synonym (e.g., 'muscle strain' vs 'myofascial injury') may be scored as incorrect despite clinical equivalence. Third, 498 of 971 ground truth records could not be definitively matched to patient IDs and were excluded from evaluation, potentially introducing selection

bias. This exclusion rate of 51.3% is substantial and may systematically favor records with complete demographic profiles, potentially biasing results toward patients with more regular visit patterns. Future work should implement a more robust probabilistic record linkage strategy (e.g., Fellegi-Sunter or Splink framework) to reduce matching failures and selection bias risk. Fourth, Typhoon2-3.1 was evaluated in its default configuration; fine-tuning or prompt optimization may substantially improve its structured extraction performance.

D. Typhoon2-3.1: Language Specialization Versus Structured Extraction Capability

The substantially lower performance of Typhoon2-3.1 (F1 = 32.9%) relative to all general-purpose comparators demands mechanistic analysis beyond surface-level format mismatch attribution. Three interdependent factors account for this divergence. First, domain-language specialization does not confer schema-constrained extraction capability. Typhoon2-3.1 was instruction-tuned on Thai conversational and task-oriented corpora to optimize fluency and cultural appropriateness in open-ended generation — an objective fundamentally distinct from strict adherence to a structured JSON output schema. This training direction appears to reinforce a strong prior toward natural language continuation, which competes with and frequently overrides the schema constraint even when the extraction instruction is explicitly provided in the prompt.

Second, the bilingual composition of the EMR records creates a specific failure mode for Thai-optimized models. Sports medicine records in this dataset mix Thai-language clinical descriptions with English anatomical terminology, numeric parameters, and institutional abbreviations within the same field (e.g., 'ปวดกล้ามเนื้อ hamstring, tenderness grade 2/5'). A model whose pre-training distribution is predominantly Thai-language assigns higher probability mass to Thai continuations, resulting in field values generated in Thai even when the schema and prompt explicitly specify English-formatted output. This is confirmed by the near-total dominance of Category (a) Empty/Format mismatch errors (97% of Typhoon2-3.1's low-F1 records), indicating that the model is not producing incorrect clinical content — it is producing non-schema-compliant output format, which is then evaluated as empty by the scoring pipeline.

Third, chat-format instruction following in Typhoon2-3.1 prioritizes conversational coherence over output structure. Under the split system/user prompt configuration used for chat-tuned models, Typhoon2-3.1 systematically produced natural language paragraph responses rather than the compact JSON structure required. This behavior was not observed in any of the four general-purpose models under identical prompting conditions. The implication for model selection is substantive: in bilingual schema-constrained extraction tasks, instruction-following fidelity and output format adherence are more critical performance determinants than language-specific pre-training. Practitioners evaluating regional language models for structured clinical NLP should assess schema compliance as a primary selection criterion, independent of language benchmark scores.

E. Inter-Annotator Agreement

Inter-annotator agreement (IAA) was assessed using Cohen's Kappa on 100 stratified records, comparing a physiotherapist (Annotator 1) using institutional shorthand and Thai language, against an AI expert (Annotator 2) using standardized English terminology. Results show that Treatment achieved the highest agreement (κ = 0.43, moderate) after normalization of abbreviations. Diagnosis (κ = -0.04) showed near-zero agreement due to non-standardized institutional shorthand codes. Chief Complaint (κ = -5.25) showed very low agreement due to Thai-English language mismatch — a formatting difference rather than clinical disagreement. These findings underscore the need for standardized terminology (SNOMED CT or ICD-11) and unified annotation language in future clinical NLP protocols, as Table V.

TABLE V. INTER-ANNOTATOR AGREEMENT: ANNOTATOR 1 (HUMAN CLINICIAN) VS. ANNOTATOR 2 (AI EXPERT), N = 100 RECORDS

Clinical Field	Ann1 (Human)	Ann2 (AI)	Mean Jaccard (Normalized)	Cohen's K (Threshold=0.5)	Interpretation
Diagnosis	Institu-tional shorthand codes (KL3, TMZ, KLV)	Standard English: MCL Grade III Sprain; Quad-ricep Strain	0.482	-0.04	Below Chance (Shorthand mismatch)
Treatment	Clinical abbreviations (ESWT, DR, ICT, HP)	Full English: Shock Wave Therapy; Dry Needling; Hot Pack	0.742	0.43 [Moderate]	Moderate (Best agreement)
Chief Complaint	Thai language text (e.g., 'ปวดหัวเข่า')	English transla-tion: Knee pain; Hip pain	0.116	-5.25	Poor (Language mismatch)
Objective Examina-tion	Mixed Thai/shorthand + partial findings	Structured English: ROM, MMT, tenderness findings	0.497	-0.19	Poor (Terminology divergence)

F. Research Contributions

This study provides three principal contributions to clinical NLP and healthcare informatics. First, we provide the first systematic benchmark of five locally-deployed open-source LLMs (Mistral-7B, Qwen2.5-7B, Gemma2-9B, LLaMA3.1-8B, Typhoon2-3.1) on Thai bilingual sports medicine EMRs, validated with Bootstrap 95% CI. This benchmark provides empirically grounded model selection guidance for privacy-preserving clinical NLP in low-resource language settings. Second, we contribute a reproducible extraction pipeline that transforms 523 unstructured Thai-English patient records into structured, machine-readable JSON format without cloud API dependency — enabling injury surveillance, treatment protocol retrieval, and longitudinal outcome analysis. Third, we generate three key empirical insights: 1) language specialization does not confer schema-constrained extraction capability — instruction-following fidelity is the dominant determinant; 2) bilingual code-switching density is a key stratification variable for multilingual clinical NLP; and 3) inter-annotator reliability is fundamentally limited by non-standardized institutional shorthand, necessitating ICD-11 or SNOMED CT alignment.

This study demonstrated that locally-deployed 7B-parameter open-source LLMs — particularly Mistral-7B and Qwen2.5-7B — achieve robust structured clinical data extraction from Thai bilingual sports medicine EMRs (F1 > 91%) without cloud infrastructure or patient data exposure. The extraction pipeline produces machine-readable JSON output from 523 patient visit records, enabling evidence-based clinical decision support, including injury surveillance, treatment protocol retrieval, and longitudinal outcome analysis. Annotation reproducibility remains limited by non-standardized institutional shorthand and Thai-English language divergence; adoption of ICD-11 or SNOMED CT terminology is recommended for future protocols. Language specialization does not confer a structured extraction advantage — instruction-following fidelity is the dominant performance determinant in bilingual clinical NLP settings.

VI. Conclusions

This study demonstrates that locally-deployed open-source 7B-parameter LLMs achieve clinically useful structured extraction accuracy (F1 > 91%) on bilingual Thai sports medicine EMRs without reliance on cloud infrastructure or third-party API services. Mistral-7B and Qwen2.5-7B produced statistically equivalent performance ($p \geq 0.10$), establishing both as viable candidates for privacy-preserving clinical NLP deployment. The finding that Thai-language model specialization does not confer advantage in schema-constrained bilingual extraction — and that instruction-following fidelity is the dominant performance determinant — provides actionable guidance for model selection in multilingual healthcare informatics. Annotation standardization through controlled clinical vocabulary aligned with ICD-11/SNOMED CT remains a prerequisite for improving inter-annotator reproducibility and scaling ground truth creation. The structured JSON output produced by this pipeline establishes a queryable clinical data layer enabling evidence-based decision support, including injury surveillance, treatment protocol retrieval, and longitudinal outcome analysis — demonstrating the translational value of local LLM deployment in resource-constrained institutional healthcare settings.

Limitations of this study include: 1) inter-annotator agreement remains constrained by non-standardized institutional shorthand and Thai-English language divergence, yielding near-zero Cohen's Kappa for the Diagnosis field ($\kappa = -0.04$) and Chief Complaint field ($\kappa = -5.25$), indicating that ground truth reliability requires future validation with trained bilingual clinical annotators and standardized vocabulary; and 2) the evaluation dataset was limited to 444 of 971 available records (45.6% utilization) due to unresolvable patient ID matching, introducing potential selection bias that may affect the generalizability of the reported performance metrics. Future work should prioritize fine-tuning selected models on ICD-11- and SNOMED CT-aligned annotated corpora to improve both extraction accuracy and annotation consistency in multilingual Thai clinical settings.

Acknowledgment

This research was supported by Thammasat University Fundamental Fund, fiscal year 2026, as allocated by Thailand Science Research and Innovation (TSRI).

References

[1] K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172-180, 2023.

[2] H. Nori et al., "Capabilities of GPT-4 on medical challenge problems," arXiv preprint arXiv:2303.13375, 2023.

[3] Office of the PDPC, "Personal Data Protection Act B.E. 2562 (2019)," Thailand, 2019.

[4] M. Agrawal et al., "Large language models are few-shot clinical information extractors," in Proc. EMNLP, 2022.

[5] D. Kartchner, S. Ramalingam, I. Al-Hussaini, O. Kronick, and C. Mitchell, "Zero-Shot Information Extraction for Clinical Meta-Analysis using Large Language Models," in Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, Canada, Jul. 2023, pp. 396–405. doi: 10.18653/v1/2023.bionlp-1.37.

[6] Ollama, "Run large language models locally," https://ollama.ai, 2024.

[7] J. Chen et al., "Feasibility of local LLM deployment for hospital information systems," J. Am. Med. Inform. Assoc., vol. 30, no. 9, pp. 1520-1528, 2022.

[8] V. Ntinopoulos et al., "Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation," BMJ Health Care Inform., vol. 32, no. 1, e101139, 2025, doi: 10.1136/bmjhci-2024-101139.

[9] J. Huang et al., "A critical assessment of using ChatGPT for extracting structured data from clinical notes," npj Digit. Med., vol. 7, p. 106, 2024, doi: 10.1038/s41746-024-01079-8.

[10] I. C. Wiest et al., "Privacy-preserving large language models for structured medical information retrieval," npj Digit. Med., vol. 7, p. 257, 2024, doi: 10.1038/s41746-024-01233-2.

[11] L. Builtjes, J. Bosma, M. Prokop, B. van Ginneken, and A. Hering, "Leveraging open-source large language models for clinical information extraction in resource-constrained settings," JAMIA Open, vol. 8, no. 5, p. ooaf109, 2025, doi: 10.1093/jamiaopen/ooaf109.

[12] M. C. Durango, E. A. Torres-Silva, and A. Orozco-Duque, "Named entity recognition in electronic health records: a methodological review," Healthc. Inform. Res., vol. 29, no. 4, pp. 286-300, 2023, doi: 10.4258/hir.2023.29.4.286.

[13] A. Kaminska et al., "Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review," Comput. Biol. Med., vol. 155, p. 106649, 2023, doi: 10.1016/j.compbiomed.2023.106649.

[14] A. Vaid, I. Landi, G. Nadkarni, and I. Nabeel, "Using fine-tuned large language models to parse clinical notes in musculoskeletal pain disorders," Lancet Digit. Health, vol. 5, no. 12, pp. e855-e864, 2023, doi: 10.1016/S2589-7500(23)00202-9.

[15] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," Proc. Natl. Acad. Sci. U.S.A., vol. 120, no. 30, e2305016120, 2023, doi: 10.1073/pnas.2305016120.

[16] K.-H. Jung, "Large language models in medicine: clinical applications, technical challenges, and ethical considerations," Healthc. Inform. Res., vol. 31, no. 2, pp. 114-124, 2025, doi: 10.4258/hir.2025.31.2.114.

[17] K. H. Lee et al., "ANNO: a general annotation tool for bilingual clinical note information extraction," Healthc. Inform. Res., vol. 28, no. 1, pp. 89-94, 2022, doi: 10.4258/hir.2022.28.1.89.

Two-Phase Transfer Learning Framework for Automated Depression Classification in the Elderly via Facial Expression Recognition

Muhammad Daffa Zahrandika Wibisono[1], Marizuana Mat Daud[2]*, Wan Mimi Diyana Wan Zaki[3]
Institute of Visual Informatics, National University of Malaysia, 43600, Bangi, Selangor Darul Ehsan, Malaysia [1, 2]
Faculty of Engineering & Built Environment, National University of Malaysia, 43600, Bangi, Selangor Darul Ehsan, Malaysia [3]

***Abstract*—Automatic detection of depression in the elderly through Facial Expression Recognition faces a fundamental challenge in the form of domain shift due to skin deformation and facial structural changes due to aging, such as ptosis and deep wrinkles. This study proposes a Two-Phase Transfer Learning framework that integrates high-density facial landmark point extraction (468 points using MediaPipe) with a hybrid spatiotemporal CNN-BiLSTM-VGG19 architecture to address these challenges. Phase I training was conducted on a standard facial dataset to obtain fundamental feature representations, followed by a fine-tuning process in Phase II using a geriatric facial dataset. Experimental results show that the CNN-BiLSTM-VGG19 architecture is highly robust, exploiting deep facial wrinkles as informative texture features. The model successfully achieved 91.42% accuracy on 70-year-old older adults. Furthermore, hyperparameter evaluation confirmed that the Stochastic Gradient Descent (SGD) optimizer combined with a low learning rate of 0.0005 was the most optimal configuration. This balance effectively prevented catastrophic forgetting during domain adaptation, while also achieving a clinical sensitivity recall rate above 96%. Comprehensively, this study demonstrates that the texture-biased CNN-BiLSTM-VGG19 model offers a robust, non-invasive, and highly efficient depression screening instrument for implementation in elderly care facilities.**

***Keywords*—Elderly depression; Facial Expression Recognition; transfer learning; VGG19; texture bias; spatiotemporal network**

I. Introduction

The phenomenon of population aging in Southeast Asia has become a serious concern, with the elderly population growing by 41.5% in the past two decades [1]. This trend aligns with conditions in Malaysia, where the population growth rate for those aged 65 and over has reached 6.1%. This significant demographic shift is correlated with an increase in mental illness, particularly depression, with a prevalence of 19.1% among the elderly [2]. A major challenge in current medical management is the high number of undetected cases [3], often caused by a decline in natural emotional expression in the elderly that masks clinical signs [4].

Conventional screening methods such as the PHQ-9 and GDS-15 questionnaires, although commonly used, are susceptible to subjective self-report bias [5]. To overcome this limitation, a facial analysis based approach is crucial, as Major Depressive Disorder (MDD) fundamentally alters the dynamics of an individual's emotional valence. The affective computing literature confirms that depression triggers negative affect potentiation, where patients tend to display predominantly sad and angry expressions, and Positive Affect Attenuation, or anhedonia, which limits the patient's ability to express happiness [6]. Therefore, Facial Expression Recognition (FER) technology does not simply perform a binary classification, but rather estimates the probability of depression based on a shift in patterns from relaxed, positive, and neutral expressions to predominantly negative affect.

More specifically, this classification analysis relies on the interpretation of different facial features between healthy and depressed groups. Referring to the findings of Girard [6], the Non Depression classification is strongly influenced by the detection of authentic happiness signals and dynamic facial muscle engagement; the absence of these signals statistically increases the predicted probability of depression due to anhedonia mechanisms. Conversely, for the 'Depression' class classification in the geriatric population, it emphasized that visual indicators are not limited to melancholic sadness but often manifest as irritability or agitation that mimics anger [6]. Therefore, an FER system must be able to integrate the dominant anger features in older adults as predictors of depression, rather than simply transient anger.

However, the application of precision FER technology to older adults faces significant technical gaps, particularly related to data availability. Existing public datasets are predominantly young adult faces and validated in controlled environments [7]. Consequently, artificial intelligence models trained on standard datasets such as CASME II or CK+ often fail to distinguish clinical emotional features from natural aging features. Permanent facial wrinkles, for example, are frequently misinterpreted as negative expressions.

Based on these challenges, this study proposes a two-phase transfer learning framework to classify depression from Non Depression in older adults. Phase I aims to extract basic emotional features from an extensive young adult dataset to build a foundation for expression recognition. Subsequently, Phase II refines the model's sensitivity through domain adaptation on an older adult facial dataset, ensuring the system can accurately distinguish between signs of pathological depression and characteristics of biological aging.

*Corresponding author

II. Literature Review

A. FER Using Deep Learning

Deep Learning represents a cutting-edge computational paradigm in artificial intelligence that models high-level abstractions on data using multi-layered processing architectures, or deep architectures [8]. Unlike traditional machine learning methods that rely heavily on hand-crafted features, Deep Learning's fundamental strength lies in its ability to perform end-to-end representation learning [9], where the model automatically discovers intrusive structures and non-linear patterns directly from raw data without human intervention [10].

In the context of Facial Expression Recognition (FER), this capability is crucial because the human face has extreme visual variability [11]. Deep Learning is able to capture subtle nuances of facial deformation and is invariant to changes in lighting and pose [12], making it a state-of-the-art standard for deciphering the complexity of human emotions that are difficult to quantify by conventional algorithms [13].

B. Spatio-Temporal and Sequential Model for FER

The development of spatio-temporal and sequential models emerged as a response to the limitations of conventional CNNs that only process spatial information on static images, where the integration of recurrent architectures such as Long Short-Term Memory (LSTM), Bidirectional LSTM, or Gated Recurrent Unit with CNNs allows the system to capture the dynamics of facial expression changes as well as the temporal relationships between frames in a video sequence [14][15]. This hybrid architecture works by extracting spatial features from each frame using a CNN, which are then sequenced and processed by recurrent layers to learn long-term dependencies between frames, so that the model can distinguish emotions based on the evolution of expressions over time [16].

Spatial features extracted by CNN represent static characteristics in each frame, but temporal information obtained from LSTM, BiLSTM allows the model to understand the context of movement and expression transitions that are crucial for distinguishing similar emotions but have different dynamic patterns [17]. The combined implementation of CNN with BILSTM can combine deep visual information with temporal sequence context simultaneously, thus enabling the model to learn expression change patterns from both time directions to improve recognition accuracy [18].

III. Methodology

The proposed methodology, as it shows on Fig. 1, begins with dataset collection, followed by feature extraction to obtain crucial facial landmarks, including the pre-processing stage. In the core stage, the method applies a two-phase transfer learning strategy: the first phase involves fine-tuning the reduction model to adapt general features, while the second phase focuses on specific training according to the intended task. Finally, the model's performance is evaluated based on its ability to recognize patterns and produce accurate predictions.

A. Dataset

To build a robust classification sample against variations in facial expressions, this study prepared a comprehensive dataset consisting of a total of 10,432 facial images, divided equally into 5,216 samples for the Depression class and 5,216 for the Non Depression class. This dataset is an amalgamation of various standard data sources, including CASME II for micro-expressions, and RAVDESS, TFEID, and CK+ for macro-expressions, as described in Table I. In determining this binary labeling, we adopted protocols validated by Girard et al. [6] and Lee [19], which showed that facial emotion profiles have a strong statistical correlation with psychological states. Individuals with depression tend to exhibit Negative Affect Potentiation (predominance of sadness and anger/irritability), while Non Depression is associated with the ability to exhibit relaxed, happy, and neutral expressions. This emotional valence-based label mapping methodology aligns with the approach of [20], which demonstrated the effectiveness of Depressed and Non Depressed labels based on emotional intensity tendencies in training Deep Learning models.

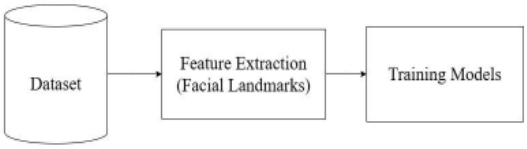


Fig. 1. Process flow of the proposed method.

This integration of multiple datasets is designed to bridge the complexity of facial dynamics, which vary from the macro to the micro level. RAVDESS plays a crucial role in capturing high-level temporal dynamics, as this dataset records not only pose statistics but also subjects speaking and singing across a wide spectrum of emotions, allowing the model to learn natural and complex expression transitions. CASME II, on the other hand, complements this analysis by capturing micro-expressions. This data presents a different challenge because expressions appear for very short durations and low intensity, yet are crucial because they represent the involuntary "emotional leaks" that are often key indicators of hidden signs of depression.

TABLE I. Technical Specifications and Dataset Sources

Dataset	Expression	Validator Institution & Key Characteristics
CASME [21]	Micro-Expression	Chinese Academy of Sciences (CAS). Recorded at high speed (200 fps) to capture involuntary micro-muscle changes.
RAVDESS [22]	Macro-Expression	Ryerson University, Canada. Provides high temporal dynamics variations through speech and song modalities.
TFEID [23]	Macro-Expression	Brain Mapping Laboratory, Taiwan. Focuses on subjects with Asian backgrounds for accurate demographic representation.
CK+ [24]	Macro-Expression	Carnegie Mellon University (CMU). The gold standard with precise FACS annotations for basic muscle movement patterns.

B. Feature Extraction

To transform raw visual information into structured data representations, this study implements a geometry-based feature extraction framework. Unlike holistic pixel-based approaches that process the entire image intensity spectrum and are highly

susceptible to environmental disturbances such as lighting fluctuations, this method focuses exclusively on the topological configuration of the face. This approach was strategically chosen to address two major challenges in facial expression analysis of elderly individuals: 1) the need to separate emotional features from skin texture and aging artifacts (such as permanent wrinkles), and 2) the urgency to capture involuntary micro movements that often go undetected in low-resolution representations.

1) High-density facial landmark detection: The crucial stage in the high-density facial landmark detection process is implemented using the MediaPipe Face Mesh Architecture. This cutting-edge framework was specifically chosen to replace conventional methods such as the sixty-eight-point Dlib Model due to its capability to reconstruct a massive three-dimensional topology consisting of four hundred and sixty-eight landmarks, as presented in Fig. 2. The fundamental justification for using this high-density topology is based on two primary technical arguments. The first argument emphasizes the increased spatial resolution essential for detecting microexpressions.

Conventional standard models have been shown to lack adequate spatial coverage of critical facial areas such as the glabella between the eyebrows and the nasolabial folds. In contrast, this new architecture provides a highly dense tessellation in these crucial areas. This level of density is essential for capturing the subtle movements of facial action units, such as action unit four, which represents the contraction of the eyebrow-lowering muscle, and action unit twelve, which represents the lip corner puller, both of which clinically serve as key indicators in identifying depressive states. The second argument relates to a high degree of pose invariance.

This framework's ability to comprehensively estimate three-dimensional spatial coordinates allows the system to consistently maintain topological consistency. This robust spatial tracking remains optimally operational even when the research subjects exhibit head pose variations up to forty-five degrees, a dynamic condition often encountered during naturalistic clinical interviews.

In the experimental setup, MediaPipe hyperparameters were strictly configured to guarantee data integrity at the frame level, as presented in Table II.

TABLE II. MEDIAPIPE HYPERPARAMETER CONFIGURATION

Parameter	Value	Description
static_image_mode	TRUE	This parameter is explicitly enabled to process each video frame as an independent entity. This approach prevents temporal bias from the tracking algorithm and ensures that landmark detection is based purely on the visual features of the current frame. This is crucial for preventing information leakage when training CNN architectures on shuffled datasets.
max_num_face	1	Limiting detection to just one face aims to focus computing resources exclusively on the primary subject. This configuration effectively eliminates background noise from other, irrelevant individuals in the frame.
min_detection_confidence	0.5	This threshold is set empirically to achieve an optimal balance between sensitivity (recall) and precision. This value ensures that the facial topology mesh will only be generated when facial features are clearly visible, significantly reducing the occurrence of false positives in ambiguous visual frames.

2) Geometry-based ROI normalization and preprocessing: After coordinate extraction, a custom preprocessing workflow is applied to normalize the Region of Interest (ROI) and eliminate external variables irrelevant to emotion, specifically lighting inconsistencies and skin tone bias. As implemented in our algorithm, this process follows three systematic steps:

a) Grayscale projection: The input image is converted to a grayscale spectrum (cv2.COLOR_BGR2GRAY). This step neutralizes chromatic lighting artifacts (e.g., the yellow tint of room lights) and removes skin color information, forcing the model to rely solely on structural deformations.

b) Structural overlay: A topological mask is drawn over a grayscale background using a Tessellation technique, connecting 468 landmarks with 1-pixel-thick, high-contrast green lines (BGR: 0, 255, 0). This effectively turns the image recognition problem into a geometric pattern recognition task.

c) Spatial standardization: The face is cropped based on the extreme coordinates of the landmarks, with 20 pixels of padding added to preserve edge detail. The resulting image is then resized to a uniform 224 x 224 pixel dimension to meet the input layer requirements.

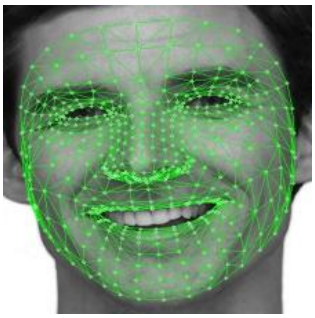


Fig. 2. Visualization of the results of feature extraction pre-processing.

A visual implementation of this framework is demonstrated in Fig. 2. The image shows the transformed result where the original color information has been removed to neutralize skin tone bias. On top of this base layer, the topological structure of the face is rendered using geometric tessellation. As seen in the details of the mouth and eye area in Fig. 2, this mesh precisely tracks facial muscle deformations such as the contraction of the Zygomaticus major during smiling, verifying that the input features passed to the classification model are pure structural movement patterns that are invariant to both subject identity and lighting conditions.

C. Training Models: Comparative Deep Transfer Learning Framework

This study applies a strategically designed Two-Phase Transfer Learning framework to build a robust depression classification model in the elderly population. The training approach is hierarchically structured to address the key challenge of labeled data scarcity in the geriatric domain.

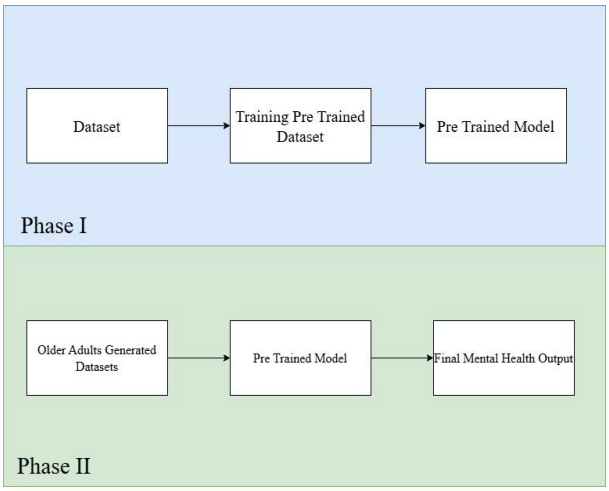


Fig. 3. Model Phase I and II.

As shown in Fig. 3, Phase I focused on Source Domain Benchmarking, where we conducted an intensive comparative study between backbone architectures such as VGG19 to learn universal spatiotemporal feature representations. Phase II focused on Target Domain Adaptation, where the best model resulting from Phase I was adapted through a fine-tuning mechanism using a synthetic elderly face dataset.

1) Phase I: Phase I served as a foundational stage, where we conducted a series of experiments to build the most optimal spatiotemporal representation model. In this stage, the model was trained from scratch to distinguish micro and macro-expression patterns associated with Depression Sad, Angry) and Non Depression (Happy, Neutral) using a standard composite dataset.

a) Dataset: To ensure robust model generalization and mitigate potential biases, Phase I of this study utilizes a Composite Source Dataset. This dataset is an amalgamation of four internationally recognized benchmark repositories, as seen in Fig. 4: Casme II (micro-expressions), Ravdess (dynamic Macro-expressions), Tfeid (Asian Facial Morphological Features), and CK+ (Action Unit Precision Validation). The integration of these diverse sources provides a comprehensive representation of facial dynamics across various temporal and morphological scales.



Fig. 4. Dataset collection from RAVDESS, CASME II, TFEID and CK+.

The acquisition and utilization of these datasets were conducted in strict adherence to ethical research protocols and privacy standards. Access to the CASME II repository was secured through a formal institutional application process. For RAVDESS, TFEID, and CK+, data were obtained under open source licenses, which involved the completion of formal consent agreements and compliance with institutional citation mandates. These datasets have been clinically and scientifically validated by their respective provider institutions, ensuring that subject privacy is preserved and that the research aligns with international ethical guidelines for automated mental health screening.

Following the amalgamation, class balancing was performed to maintain an equitable distribution between classes. The final dataset for Phase I comprises 10,432 facial image samples, symmetrically partitioned into 5,216 samples for the 'Depression' class and 5,216 for the 'Non Depression' class. For model development and empirical evaluation, a systematic 80:20 data split ratio was implemented. Consequently, 8,346 images were allocated for the training phase, while 2,086 images were strictly reserved for the testing and validation set. This explicit partition addresses the requirement for statistically significant performance evaluation on unseen data.

All samples underwent a standardized preprocessing pipeline: conversion to grayscale format, application of a high-density topological dense mesh structure (green), and rescaling to a uniform dimension of 224 x 224 pixels. This structured input format is specifically engineered to optimize compatibility with the CNN-BiLSTM-VGG19 architecture, facilitating the extraction of invariant geometric facial features while eliminating confounding color variations.

b) Trained pre-train model: The training process in Phase I was designed through a multi-stage experimental approach (multi-stage benchmarking) to thoroughly validate each architectural component.

c) Preliminary ablation study: Spatial vs. Spatiotemporal: Before finalizing the main feature extraction architecture, a preliminary ablation study comparing spatial and spatiotemporal approaches was conducted to validate the importance of integrating the temporal module. This test was designed for a direct performance comparison between a purely spatial model represented by CNN-VGG19 and a hybrid spatiotemporal model based on CNN-BiLSTM-VGG19. In the former, the model works exclusively by relying on static spatial feature extraction capabilities. In contrast, the hybrid approach integrates a Bidirectional Long Short-term Memory Layer or BiLSTM specifically designed to comprehensively capture temporal dependencies between frames. This series of initial experimental results specifically aimed to provide empirical evidence that the addition of the BiLSTM module is able to demonstrate a very significant improvement in the accuracy of microexpression recognition when compared to purely spatial-based methods.

d) Backbone architecture benchmarking: After validating the superiority of the hybrid architecture, the next step focused on selecting the most optimal primary spatial feature extractor. This experiment was specifically designed to

answer the fundamental question of whether the density of texture extraction in VGG is superior in capturing microexpressions related to Depression. To answer this question, this study compares two variants of the hybrid architecture with very contrasting characteristics. The first variant, CNN-BiLSTM-VGG19, is built on a stack of nineteen convolutional layers with a small kernel size of 3 X 3. This architectural design allows the model to learn a very detailed hierarchical feature representation ranging from simple edge extraction to complex facial texture patterns. In the context of Depression detection, VGG19 demonstrates comparative advantages in extracting very fine texture nuances such as the depth of forehead wrinkles and shadows around the eyes.

Through synergy with BiLSTM, these static textural features are transformed into a continuous time sequence, enabling the model to not only detect the presence of wrinkles but also understand the dynamics of their depth and fading over time. The hypothesis proposed for this variant is that VGG19's sensitivity to pixel-level details makes it a very strong candidate for processing low-resolution inputs and micro-feature extraction.

e) Hyperparameter grid search optimization: To ensure a comprehensive comparison, all models were trained using the rigorous grid search scheme by testing various parameter combinations to find the best convergence configuration. This testing included optimizing Adam for fast convergence and SGD for generalization stability, as well as testing learning rates across a wide spectrum of values, including 0.01, 0.05, 0.001, 0.005, 0.0005, and 0.0001. The training duration was also set at a maximum limit of 100 epochs by implementing an early stopping mechanism with a tolerance limit of 10 epochs to prevent overfitting and stop the training process if there is no improvement in the validation loss.

f) Pre-trained model: The final result of Phase I is the best-performing pre-trained model selected based on the highest validation accuracy and lowest loss metric across all combinations of experiments. This selected VGG19 model stores basic knowledge about universal facial expression dynamics within its synaptic weights. This model is then stored and will serve as an intelligent initiator for the transfer learning process in Phase II to replace the less efficient random weight initialization. Next, this VGG19 model is combined with the BiLSTM architecture to continuously optimize expression recognition. This combination is implemented because VGG19 has high reliability in extracting detailed spatial features, while BiLSTM is very effective in capturing the temporal dependencies of a series of bidirectional emotional transitions. Therefore, the combination of the two can produce a system that understands the dynamics of facial changes much more comprehensively and accurately.

g) Phase II: Phase II is the specialization stage where the general knowledge about facial expressions acquired in Phase I is transferred to the target domain, namely the elderly population. This stage is crucial to address the domain shift problem, where models trained on young faces often fail to adapt to the complex morphological characteristics of aging faces, such as deep wrinkles and decreased skin elasticity (ptosis). Instead of training the model from scratch, Phase II applies a transfer learning strategy to refine the model's sensitivity to the nuances of Geriatric faces.

h) Older adults dataset: A major challenge in analyzing depression in the elderly is the lack of public datasets containing high-quality micro- and macro-expressions in Geriatric subjects. To address this without sacrificing precise emotion labels, Phase II utilized the Synthetic Elderly Dataset, generated through a generative transformation of all source datasets from Phase I, such as Casme II, Ravdess, Tfeid, and CK+.

Algorithmically, the U-Net model in Face Re-Aging Network (FRAN) receives input in the form of a 5-channel tensor. This tensor consists of the original RGB image to be aged, as well as two single-channel age maps, each representing the input current age and the target age (in this case, set to 75 years old). The pixel values in these age maps are normalized to a range of 0 to 1 to represent a continuous age interval [25].

To ensure experimental consistency and statistical rigor, the total volume of the Phase II dataset was maintained at 10,432 facial image samples, mirroring the distribution of Phase I (5,216 samples for both Depression and Non Depression classes). This consistency allows for a direct comparative analysis of the model's performance before and after the domain adaptation to geriatric features. In line with the established methodology, a stratified 80:20 split ratio was applied: 8,346 images were utilized for fine-tuning the hybrid architecture, while 2,086 images were allocated for the terminal testing and validation phase.

A fundamental advantage of the FRAN architecture, crucial for microexpression analysis, is its ability to preserve the identity and spatial geometry of the original face. Instead of generating a completely new facial image, which is prone to identity loss, the U-Net model predicts per-pixel RGB deltas or offsets. These aging deltas are then added to the original RGB input image to produce the final aged image. This mechanism, combined with the U-Net's skip connections, ensures that the subject's high-resolution facial features and emotional expressions are not distorted. FRAN has proven robust in adding realistic aging textures, such as deep wrinkles, sagging skin (ptosis), and changes in cartilage proportions, such as in the ear and nose areas, while maintaining temporal consistency, head pose, and lighting variations.

This approach effectively produces geriatric twins that retain the original emotional semantics of young subjects while capturing realistic and challenging facial aging features. These generative transformation specifications are strategically applied to various datasets to meet comprehensive testing objectives. In the Aged-CASME II and CK+ datasets, the transformations are specifically used to train a classification model to recognize and distinguish the complex interactions between static wrinkles inherent in older adults and dynamic wrinkles triggered by microexpressions. Modifications to the Aged-RAVDESS dataset aim to simulate older subjects with high expression intensity or who are speaking, a representation highly relevant for supporting video analysis of real-world clinical interviews. Furthermore, the transformations in the Aged-TFEID dataset play a crucial role in generating synthetic

data of Asian older adults aligned with the study's target demographic, significantly reducing racial bias in the model's final inference capabilities.

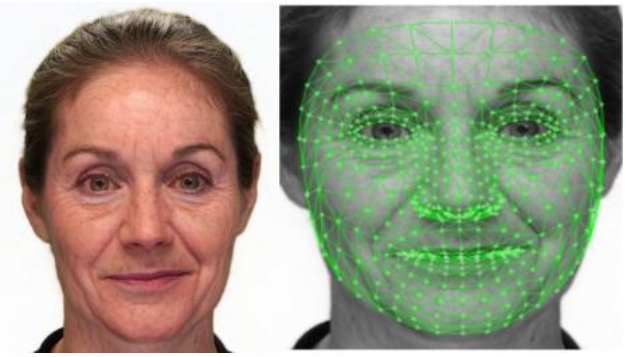


Fig. 5. Left: [Fig. 5(a)] - Synthetic elderly face in color | Right: [Fig. 5(b)] - Grayscale face with green mesh.

To visualize the results of the generative transformation and data preparation process, Fig. 5 presents a composite example from a sample of the Synthetic Elderly Dataset. Fig. 5(a) displays the raw output of the FRAN U-Net model with a target age of 75, clearly demonstrating the addition of aging features such as deep nasolabial folds and sagging skin around the eyes. Fig. 5(b) shows the same image after the standard preprocessing steps described in Section III-B, where the color information has been normalized to grayscale and a high-density MediaPipe mesh topology is overlaid on the face. These geometric and textural representations in Fig. 5(b) serve as the actual input to the CNN-BiLSTM model during the domain adaptation of the fine-tuning phase.

2) Pre-trained model CNN-BiLSTM: The training mechanism in Phase II applies a transfer learning via a fine-tuning strategy. We reload the weights from the best-performing pre-trained model whose performance has been validated in Phase I. This training strategy is designed to prevent damage to learned spatial features (Feature destruction) through the following steps:

a) Partial freezing: The initial convolutional layers (lower layers) in the backbone are frozen. These layers are responsible for detecting basic features such as edges and geometric shapes that are universal and do not change significantly with aging. Fine-tuning high-level layers: the higher-level convolutional layers and the entire temporal module (BiLSTM) are unfrozen. This allows the model weights to adapt specifically to the skin texture of older adults and the dynamics of movement that may be slower or stiffer.

b) Learning rate: The learning rate is significantly reduced compared to Phase I. This reduction aims to prevent catastrophic forgetting, ensuring that the model only makes subtle weight adjustments without forgetting basic knowledge of facial expressions.

Similar to Phase I, we again compared the performance of the VGG19architectures on this synthetic dataset. The goal was to test the robustness of each architecture's adaptation regarding VGG's texture features that were more robust in handling visual noise caused by aging.

3) Mental health output: The final result of Phase II is the Specialized Elderly Depression Classifier. This model is an evolution of the Phase I model, which already had a specific "intuition" for elderly faces. The output layer uses a Softmax activation function that generates probability distributions for two diagnostic classes: Depression and Non Depression.

The final classification is based on an interpretation of emotional valence validated by the clinical literature [26][27]:

a) Depression: Classified based on the probability dominance of persistent negative expression features, particularly Sadness and Anger. This reflects the phenomenon of negative affect potentiation in people with major depression, where subjects tend to maintain negative affect for longer.

b) Non Depression: Classified based on the stability of a relaxed neutral expression and the presence of the happiness feature. The dominance of this feature reflects the absence of negative interpretation bias and healthy emotional regulation function.

This final model is ready to be used for inference on real clinical data as a non-invasive and objective initial screening tool for mental health in older adults.

IV. Results and Discussion

This section presents a comprehensive analysis of the performance of a Deep Learning architecture developed for depression detection in the elderly population using a Two-Phase Transfer Learning scheme. Experimental evaluations are systematically designed to address the research questions, starting from the development of a baseline model in the source domain to testing the model's robustness in the target domain, which has complex aging features.

A. Phase I: Source Domain Benchmarking (Training and Classification)

The first phase of this research focused on building a robust model baseline. The model was trained using a Composite Dataset encompassing a variety of universal micro and macro expressions. The primary goal of this phase was to identify the optimal combination of optimizer and learning rate that achieves optimal convergence, which will serve as the smart initialization for the adaptation process in Phase II.

1) Phase I: Classification results (Testing Metrics): A comprehensive quantitative evaluation was conducted on 12 different experimental scenarios on the testing set of the source domain. Table III, below, presents a complete summary of model performance based on hyperparameter variations.

Best Model Identification: Based on the comprehensive data in Table III, the SGD configuration with a Learning Rate of 0.0005 in No 8 was determined as the best model. This model dominates with an Accuracy of 96.21% and an F1-Score of 0.9621, demonstrating perfect consistency in recognizing the Depression and Non Depression classes, and significantly outperforming all Adam variations.

TABLE III. COMPLETE RECAPITULATION OF PHASE I CLASSIFICATION RESULTS (CNN-BILSTM-VGG19).

No	Optimizer	Learning Rate	Accuracy	Precision	Recall
1	Adam	0.0001	0.7411	0.8225	0.7411
2	Adam	0.0005	0.686	0.7923	0.686
3	Adam	0.001	0.7891	0.8431	0.7891
4	Adam	0.005	0.8337	0.8578	0.8337
5	Adam	0.01	0.7301	0.7447	0.7301
6	Adam	0.05	0.5	0.25	0.5
7	SGD	0.0001	0.5427	0.6295	0.5427
8	SGD	0.0005	0.9621	0.9628	0.9621
9	SGD	0.001	0.908	0.9178	0.908
10	SGD	0.005	0.93	0.935	0.93
11	SGD	0.01	0.8509	0.8747	0.8509
12	SGD	0.05	0.9521	0.9536	0.9521

2) Training dynamics analysis: The varying classification results in Table III above were further analyzed by reviewing the loss and accuracy dynamics during the training process. This analysis revealed the scientific rationale for the superiority of the combination of SGD and LR 0.0005.

A comparative analysis between the Stochastic Gradient Descent (SGD) and Adam optimizers reveals significant differences in stability characteristics. Based on the first through sixth experiments, the Adam optimizer tends to exhibit instability on this dataset. Although Adam is known for its fast initial convergence rate, this method often gets stuck in sharp minima (local minima), resulting in poor model generalization. For example, using a Learning Rate of 0.001, despite high training accuracy, the testing accuracy only reached 78.91%.

In contrast, the SGD optimizer, particularly in experiments 8, 9, 10, and 12, consistently demonstrated high performance with accuracy above 90%. SGD's more gradual weight update mechanism allows it to find flat minima (sloping minimum points). Convergence solutions located in these gentle areas are more robust, allowing the model to maintain its accuracy even when faced with new variations in the testing dataset.

In addition to optimizer selection, analysis of the Learning Rate precision demonstrates that this parameter is a key determinant of model success or failure. In the case of underfitting with a learning rate that is too small, namely Learning Rate < 0.0005, as in experiment 7, namely the CNN-BILSTM SGD Model 0.0001, the model accuracy dropped to 45.06%. This occurs because the model fails to learn complex features within the specified limit of 100 epochs. On the other hand, the overshooting phenomenon occurs when the learning rate is too large, such as Learning Rate > 0.01, as seen in experiment 6, namely the model with the Adam optimizer 0.05 and the model with the SGD optimizer 0.01, which experienced a drastic decrease in performance.

An overly aggressive learning rate causes the gradient step to "jump" the optimal point, namely the lowest loss point, and ultimately damages the model's weight structure. Therefore, the LR value of 0.0005 in the SGD Model in Experiment 8 is the perfect balance point (sweet spot). This value proved to be large enough to help the model escape local minima, but still small enough to perform precise fine-tuning, as evidenced by the lowest Validation Loss achieved with a value of 0.2419 among all configurations.

B. Phase II: Target Domain Adaptation (Augmented Elderly Dataset)

After successfully building a robust baseline model in Phase I using a Universal Dataset, the research proceeded to the Transfer Learning stage to adapt the model to a more specific target domain. In this phase, the optimal weights (pre-trained weights) from the CNN-BiLSTM-VGG19 model trained in Phase I were reloaded and fine-tuned.

The dataset used in this stage was the Augmented Elderly Dataset, which represents facial images of 70-year-olds. The purpose of this augmented data (U-Net generation) was to train the model to recognize emotional features amidst visual noise typical of aging, such as forehead wrinkles, drooping eyelids (ptosis), and sagging skin texture, while maintaining the previously learned baseline feature representation.

1) Phase II: Classification results (Testing Metrics): A comprehensive evaluation of the model's adaptation performance was conducted using a 70-year-old elderly person as the test set. Table IV below presents a complete summary of the results of fine-tuning experiments with various optimizers and learning rate variations to determine the best adaptation configuration.

TABLE IV. COMPLETE RECAPITULATION OF PHASE II CLASSIFICATION RESULTS (DATASET OF ELDERLY PEOPLE 70 YEARS AND OVER - CNN-BILSTM-VGG-19).

No	Optimizer	Learning Rate	Accuracy	Precision	Recall
1	Adam	0.0001	0.7411	0.8225	0.7411
2	Adam	0.0005	0.686	0.7923	0.686
3	Adam	0.001	0.7891	0.8431	0.7891
4	Adam	0.005	0.8337	0.8578	0.8337
5	Adam	0.01	0.7301	0.7447	0.7301
6	Adam	0.05	0.5	0.25	0.5
7	SGD	0.0001	0.5427	0.6295	0.5427
8	SGD	0.0005	0.9621	0.9628	0.9621
9	SGD	0.001	0.908	0.9178	0.908
10	SGD	0.005	0.93	0.935	0.93
11	SGD	0.01	0.8509	0.8747	0.8509
12	SGD	0.05	0.9521	0.9536	0.9521

C. In-Depth Analysis: Model Adaptation to Elderly Faces

Based on the quantitative data in Table IV, the Transfer Learning process from Phase I to Phase II exhibits complex dynamics. The selection of the best model is not based solely on accuracy but also considers the sensitivity metric (Recall), which is crucial for medical diagnosis.

In determining the best model in Phase II, although the SGD configuration with a learning rate of 0.001 in Experiment 9

achieved the highest marginal accuracy of 92.19 per cent, this study determined SGD 0.0005 in Experiment 8 as the most optimal final model. This decision was based on clinical considerations prioritizing recall or sensitivity metrics.

The SGD 0.0005 configuration produced the highest recall value of 0.9271, outperforming the SGD 0.001 configuration at 0.9080. In the context of a medical decision support system for depression detection, false negatives, or the system's failure to detect a patient who is actually ill, carry a much higher risk of fatality than false positives. Therefore, a model with high positive detection capability is an absolute priority. Furthermore, with an F1-score of 0.9153, the SGD 0.0005 configuration demonstrates a harmonious balance between accuracy and sensitivity. This demonstrates that the model is capable of adapting to the wrinkled texture of elderly faces without losing its discriminatory ability.

The success of the 0.0005 SGD model in achieving accuracy above 91 per cent also validates the effectiveness of the transfer learning strategy, particularly when fine-tuning with a low learning rate. Using a learning rate of 0.0005 allows the model to make micro-adjustments to its convolutional weights. Through this mechanism, the model relearns to interpret aging lines on elderly faces as representing texture features, rather than simply visual noise, while retaining the structural facial knowledge learned in Phase I. Conversely, using too high a learning rate proved highly destructive, as evidenced by the catastrophic forgetting phenomenon in Experiment 11 with an SGD setting of 0.01. In this configuration, the model's accuracy plummeted to 50.00 percent, equivalent to the probability of a random guess in binary classification. This complete failure indicates that the overly aggressive gradient update step during the fine-tuning process damaged the optimal pre-trained weight structure from Phase I. As a result, the model lost all memory of facial feature representations and failed to recognize any patterns in the elderly data set.

Furthermore, this experiment also revealed the limitations of the Adam optimizer when faced with domain shift phenomena. Overall, testing with the Adam optimizer from Experiments 1 to 6 yielded significantly lower recall performance than SGD. As a representative example, the Adam configuration of 0.0005 in Experiment 2 achieved a fairly good accuracy of 89.07 percent with a high precision of 93.77 percent, but its recall value remained at 0.8370. This finding indicates that Adam tends to be conservative and often predicts the majority class, namely the Non Depressed class, to avoid prediction errors. This behavior ultimately makes the model less sensitive to capturing subtle manifestations of depressive emotions behind facial skin deformation in the elderly.

Concluding Phase II, this comprehensive analysis confirmed that the CNN-BiLSTM-VGG19 architecture combined with the SGD optimizer and a learning rate of 0.0005 is the most robust configuration for transferring expression detection knowledge to the geriatric domain. This model has successfully overcome the challenges of visual complexity in facial data augmentation of 70-year-old elderly with excellent sensitivity, and is certainly ready to undergo robustness testing on more extreme data sets in the next stage.

V. Conclusion

Based on the experimental results and analysis, this study concludes that the CNN-BiLSTM-VGG19 architecture is proven effective in handling domain shift in elderly faces, with an accuracy of 91.42% and a recall of 0.9271 on the 70-year-old dataset, making it a safe screening tool with minimal risk of false negatives. The model's performance is supported by an optimal configuration through a fine-tuning strategy using the SGD optimizer with a learning rate of 0.0005, a setting that has proven to be the most stable in preventing catastrophic forgetting while providing better generalization compared to the Adam optimizer or the use of high learning rates that are prone to triggering model collapse.

References

[1] ASEANstats. (2023). Ageing ASEAN: Shifting Demographic Structure [Infographic].

[2] Jalali, A., Ziapour, A., Karimi, Z., Rezaei, M., Emami, B., Kalhori, R. P., Khosravi, F., Sameni, J. S., & Kazeminia, M. (2024). Global prevalence of depression, anxiety, and stress in the elderly population: a systematic review and meta-analysis. BMC Geriatrics, 24(1), 809.

[3] Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9 Validity of a Brief Depression Severity Measure. Journal of General Internal Medicine, 16(9), 606-613.

[4] Cui, L., Li, S., Wang, S., Wu, X., Liu, Y., Yu, W., Wang, Y.,

[5] Tang, Y., Xia, M., & Li, B. (2024). Major depressive disorder: hypothesis, mechanism, prevention and treatment. Signal Transduction and Targeted Therapy, 9(1).

[6] Parikh, A., Sadeghi, M., Richer, R., Rupp, L. H., Schindler-Gmelch, L., Keinert, M., Hager, M., Capito, K., Rahimi, F., Egger, B., Berking, M., & Eskofier, B. M. (2024). <i>Exploring Facial Biomarkers for Depression through Temporal Analysis of Action Units

[7] Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S., & Rosenwald, D. P. (2013). Social risk and depression: Evidence from manual and automatic facial expression analysis. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013).

[8] Pot, A., Carstensen, L.L. A generated image repository of aging faces. Sci Data 12, 1610 (2025). https://doi.org/10.1038/s41597-025-05909-6.

[9] Ekundayo, O. S., & Ezugwu, A. E. (2025). Deep learning: Historical overview from inception to actualization, models, applications and future trends. Applied Soft Computing, 181, 113378. https://doi.org/10.1016/j.asoc.2025.113378

[10] Mumuni, A., & Mumuni, F. (2025). Automated data processing and feature engineering for deep learning and big data applications: A survey. Journal of Information and Intelligence, 3(2), 113–153. https://doi.org/10.1016/j.jiixd.2024.01.002

[11] Sarker IH. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Comput Sci. 2021;2(6):420. doi: 10.1007/s42979-021-00815-1. Epub 2021 Aug 18. PMID: 34426802; PMCID: PMC8372231.

[12] Singh, S., & Prasad, S. V. A. V. (2018). Techniques and challenges of face recognition: A critical review. Procedia Computer Science, 143, 536–543. https://doi.org/10.1016/j.procs.2018.10.427

[13] Zhalgas, A., Amirgaliyev, B., & Sovet, A. (2025). Robust Face Recognition Under Challenging Conditions: A Comprehensive Review of Deep Learning Methods and Challenges. Applied Sciences, 15(17), 9390. https://doi.org/10.3390/app15179390

[14] Yazici, A., Kucukyilmaz, T., Dokeroglu, T., Sharipbay, A., Lee, M.-H., & Tyler, B. (2026). State-of-the-art multimodal emotion recognition: A comprehensive survey and taxonomy. Intelligent Systems with Applications, 30, 200642. https://doi.org/10.1016/j.iswa.2026.200642

[15] Gao, Y., Xiao, Z., Gong, Z., Huang, S., & Zhu, H. (2025). Spatiotemporal Deformation Prediction Model for Retaining Structures Integrating ConvGRU and Cross-Attention Mechanism. Buildings, 15(14), 2537. https://doi.org/10.3390/buildings15142537

[16] Hayat MT, Allawi YM, Alamro W, Sultan SM, Abadleh A, Kang H, Zreikat AI. A Hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM)-Attention Model Architecture for Precise Medical Image Analysis and Disease Diagnosis. Diagnostics (Basel). 2025 Oct 23;15(21):2673. doi: 10.3390/diagnostics15212673. PMID: 41225966; PMCID: PMC12608630.

[17] Amadi, Christian & Odii, Juliet & Ofoegbu, Christopher & Okpalla, Chidimma. (2023). Emotion Detection Using a Bidirectional Long-Short Term Memory (BiLSTM) Neural Network. International Journal of Current Pharmaceutical Review and Research. Vol 4, no 11. 1718-1732.

[18] Yadav, Srijana & Mangalampalli, S.. (2025). Deepfake defense: Combining spatial and temporal cues with CNN–BiLSTM–transformer architecture. PLOS One. 20. 10.1371/journal.pone.0334980.

[19] Najia, Mechichi & Faouzi, Benzarti. (2025). An Enhanced Hybrid Model Combining CNN, BiLSTM, and Attention Mechanism for ECG Segment Classification. Biomedical Engineering and Computational Biology. 16. 10.1177/11795972251341051

[20] Lee, T., Baek, S., Lee, J., Chung, E. S., Yun, K., Kim, T. S., & Oh, J. (2024). A Deep Learning Driven Simulation Analysis of the Emotional Profiles of Depression Based on Facial Expression Dynamics. Clinical Psychopharmacology and Neuroscience, 22(1), 87–94.

[21] Sugiyanto, S., Purnama, I. K. E., Yuniarno, E. M., Anggraeni, W., & Purnomo, M. H. (2024). Depression Classification Based on Facial Action Unit Intensity Features Using CNN-Poolingless Framework. International Journal of Intelligent Engineering and Systems, 17(5), 172–187.

[22] Yan, W.-J., Li, X., Wang, S.-J., Zhao, G., Liu, Y.-J., Chen, Y.-H., & Fu, X. (2014). CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PLOS ONE, 9(1), Article e86041. https://doi.org/10.1371/journal.pone.0086041

[23] Livingstone, S. R., & Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A highly validated multimodal database of emotional speech and song. PLOS ONE, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

[24] Chen, J. W., & Yen, N. S. 2007. Taiwanese Facial Expression Image Database (TFEID). Brain and Consciousness Research Center, National Chengchi University.

[25] Lucey, P., et al. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 94-101

[26] Garcia, R., et al. 2022. FRAN: Face Re-aging Network for Realistic Geriatric Transformation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2415-2424.

[27] DOI : 10.1109/CVPR52688.2022.

[28] Li, M., Li, J. X., Han, J. M., Liu, X. H., Gao, X. Z., Chen, L. M., Zhou, Z. H., & Zhou, H. L. 2026. The Socio-emotional preference task in major depressive disorder: ERP evidence of social appraisal dysfunction. BMC Psychiatry, 26(31). https://doi.org/10.1186/s12888-025-07684-5

Comparison of Time-Domain and Frequency-Domain EMG Features for Gait Phases Classification Using Machine Learning

Muhamad Amirul Sunni Rohim[1], Nurhazimah Nazmi[2]*,
Shin-Ichirou Yamamoto[3], Muhammad Kashfi Shabdin[4], Mohd Asyadi Azam[5]
Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur,
Jalan Sultan Yahya Petra, 54100, Kuala Lumpur, Malaysia[1, 2]
Department of Bio-Science and Engineering, Shibaura Institute of Technology, Fukasaku 307, Saitama-City 337-8570, Japan[2, 3]
Superconductor and Thin Film Laboratory-Faculty of Science-Department of Physics,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia[4]
Division of Advanced Applied Physics, Institute of Engineering,
Tokyo University of Agriculture and Technology, 2-24-16, Naka-cho, Koganei-shi, Tokyo 184-8588, Japan[4]
Faculty of Industrial and Manufacturing Technology and Engineering,
Universiti Teknikal Malaysia Melaka, Durian Tunggal, Melaka, 76100, Malaysia[5]

***Abstract*—Accurate gait phase detection is essential for biomechanical analysis and the control of wearable assistive devices such as powered prostheses and exoskeletons. Electromyography (EMG) provides a direct representation of neuromuscular activation and offers potential advantages for low-latency, anticipatory gait phase recognition. However, the effectiveness of different EMG feature representations for stance-swing classification has not yet been clearly established. Therefore, this study presents a systematic comparison of time-domain (TD) and frequency-domain (FD) EMG features for gait phase classification. EMG signals were recorded from the tibialis anterior and medial gastrocnemius muscles of ten healthy participants during level walking. After preprocessing and segmentation, TD and FD features were extracted and used as inputs to a support vector machine classifier with a radial basis function kernel. Model performance was evaluated using a leave-one-subject-out cross-validation framework to assess generalization. The results demonstrate that TD features consistently outperform FD features across all evaluation metrics, achieving an accuracy of 0.813 ± 0.112, macro-averaged F1-score (Macro-F1) of 0.812 ± 0.114, and Matthews correlation coefficient (MCC) of 0.672 ± 0.178, compared to FD features with an accuracy of 0.712 ± 0.077, Macro-F1 of 0.708 ± 0.079, and MCC of 0.448 ± 0.159. These findings indicate that TD features more effectively capture the transient amplitude-based neuromuscular patterns associated with gait phase transitions. In addition, TD features offer lower computational complexity, making them well-suited for real-time implementation. Overall, this study highlights the superiority of time-domain EMG representations for reliable and efficient gait phase detection and provides practical guidance for the development of wearable gait monitoring and assistive control systems.**

***Keywords*—Electromyography (EMG); gait phase detection; stance–swing classification; time-domain features; frequency-domain features; support vector machine (SVM); wearable assistive devices**

I. Introduction

Human gait is a cyclic neuromuscular behavior that arises from the coordinated activation of multiple lower-limb muscles to alternately support body weight and enable forward progression. For each leg, the gait cycle is commonly described by two principal phases: stance, during which the foot is in contact with the ground, and swing, during which the foot is lifted and advanced. These phases form the fundamental state representation used in many gait analysis and control frameworks [1,2]. Accurate and low-latency identification of gait phases is essential in biomechanics and rehabilitation engineering, particularly in wearable assistive technologies where mechanical assistance must be synchronized with the user's movement intent. In this context, a comprehensive review of lower-limb exoskeleton systems highlighted that reliable intention recognition, particularly through physiological signals such as electromyography (EMG), is critical for effective operation, while also identifying persistent challenges related to system stability and user comfort [2]. Complementing this system-level perspective, recent advances in machine learning have demonstrated that gait phase recognition can be achieved with high accuracy, exceeding 94% in multi-phase classification tasks, thereby supporting the feasibility of data-driven control strategies [3]. More broadly, artificial intelligence-based approaches have been increasingly adopted for locomotion classification and control, although their clinical validation remains limited [4]. Furthermore, existing exoskeleton systems predominantly rely on rule-based and trajectory-tracking strategies, with relatively low adoption of adaptive control methods despite their potential to improve responsiveness and user-specific adaptation [5]. Collectively, these findings emphasize that robust estimation of gait state and user intent remains central to achieving safe, comfortable, and clinically effective assistance, particularly in real-world applications involving diverse users.

*Corresponding author

Wearable gait phase recognition has traditionally relied on kinematic and kinetic sensors such as inertial measurement units and pressure insoles, which are effective for detecting gait events but primarily reflect motion and contact changes after their occurrence [6,7]. Electromyography (EMG) provides a complementary perspective by directly capturing neuromuscular activation and has therefore been widely investigated as a human–machine interface for gait monitoring and assistive device control. This is attributed to the electromechanical delay, whereby muscle activation precedes observable limb kinematics by approximately 120 ms, enabling earlier insight into movement intent [8]. Similarly, a metric learning-based temporal convolutional network has achieved high classification accuracy while improving robustness to noise disturbances, highlighting the reliability of EMG signals in complex environments [9]. Additional studies have further confirmed the effectiveness of EMG-based recognition using diverse approaches, including multi-scale deep architectures, domain adaptation strategies, and multi-feature fusion methods [10–12]. This anticipatory characteristic, commonly referred to as electromechanical delay, enables EMG-based systems to reduce effective control latency and improve synchronization of assistive actions. It has also motivated the integration of EMG with complementary sensing modalities in multimodal frameworks to enhance performance [13]. Consequently, EMG has attracted increasing attention for online stance–swing recognition in wearable systems. Moreover, the structured activation patterns of lower-limb muscles during gait further support its discriminative capability; for example, tibialis anterior activation occurs before foot contact as a preparatory response to ground interaction, while plantarflexor muscles contribute to propulsion during late stance [14]. In addition, EMG-derived features have been shown to reliably reflect key gait parameters, including stance and swing phases, reinforcing the ability of neuromuscular signals to capture phase-specific characteristics of locomotion [15]. Related studies have also extended EMG-based prediction to more complex locomotor intentions such as directional stepping, further demonstrating the richness of neuromuscular information for wearable control applications [16].

Consequently, a growing body of research has applied machine learning and deep learning techniques to EMG-based gait phase recognition, particularly for real-time exoskeleton control and subject-independent stance–swing estimation. For instance, a BiLSTM-based framework has demonstrated reliable gait phase identification during exoskeleton-assisted walking, with the ability to generalize across different locomotion conditions [8]. Further optimization of deep learning architectures has shown that high classification accuracy can be achieved with significantly reduced computational cost, making such models suitable for real-time deployment [1]. In addition, a metric learning-based temporal convolutional network has been shown to improve both recognition accuracy and robustness to noise disturbances, highlighting the importance of discriminative feature learning for reliable performance in complex environments [9]. Other studies have explored alternative strategies, including domain adaptation to address gait speed variability [11], multi-feature fusion combining time-domain, frequency-domain, and nonlinear descriptors [12], and multimodal sensor fusion frameworks integrating EMG with complementary signals to enhance recognition accuracy [13]. However, despite these methodological advances, an important practical design question remains insufficiently addressed: which EMG feature-domain representation most effectively supports reliable stance–swing classification when evaluated under controlled conditions, including consistent classifier selection and subject-independent validation. Recent surveys of EMG analysis emphasize that feature representation is a major determinant of performance, robustness, and computational efficiency, typically categorizing handcrafted features into time-domain, frequency-domain, and time–frequency domains with different cost–information trade-offs [17]. Within gait phase recognition research, feature selection remains highly heterogeneous, ranging from low-cost time-domain descriptors to hybrid feature sets and end-to-end learned representations, often without systematic comparison to determine which feature domain is most appropriate for stance–swing separability under realistic deployment conditions [8,9,11–13].

The choice of feature domain is particularly important for stance–swing classification because surface EMG during walking is characterized by transient, phase-locked bursts of muscle activation and amplitude modulations that are inherently represented in the time domain [14–17]. Accordingly, time-domain descriptors directly capture the temporal structure and magnitude of neuromuscular activity associated with gait events. In contrast, frequency-domain features summarize spectral properties of the signal that are more closely related to broader physiological conditions, such as muscle fatigue, commonly quantified using measures such as median or mean frequency [18,19]. Although spectral features may provide complementary information in certain EMG analysis tasks, their contribution to gait phase discrimination is less direct, as stance–swing transitions are primarily governed by the timing and amplitude patterns of muscle activation. This is supported by studies demonstrating that anticipatory muscle activity plays a central role in real-time gait phase estimation for exoskeleton control, where temporal alignment of EMG signals enables early detection of gait events [1,8]. In addition, multimodal approaches have shown that EMG contributes most effectively through its temporal continuity and activation patterns rather than through spectral characterization alone [13]. Furthermore, real-time wearable control systems, including powered prostheses and lower-limb exoskeletons, impose strict latency and computational constraints, which necessitate feature representations that balance discriminative capability with efficiency [9,13]. In this context, recent studies on EMG-driven intent estimation highlight a practical accuracy–latency trade-off and consistently report that time-domain features offer a more favorable balance between classification performance and computational efficiency compared with frequency-domain alternatives in time-constrained control applications [20].

Therefore, a systematic and controlled comparison of time-domain versus frequency-domain EMG features for stance–swing classification is needed to clarify which representation more effectively captures the neuromuscular signatures of gait phases under subject-independent validation. In this study, lower-limb EMG recorded from TA and MGAS during walking was used to extract feature sets in the time and frequency domains. A support vector machine (SVM) classifier was used

to evaluate stance–swing classification, and performance was assessed under a subject-independent validation framework. By holding the classification model and evaluation protocol constant while varying only the feature-domain representation, the study aims to provide actionable guidance for designing robust, computationally efficient EMG-based gait phase detection pipelines for wearable gait monitoring and assistive device control.

II. Related Work

Wearable gait phase detection has traditionally relied on kinematic and kinetic sensing modalities, particularly inertial measurement units and plantar-pressure or insole systems, as these sensing approaches provide reliable detection of gait events and are well established in prosthetic and exoskeleton applications [6,21,22]. However, these signals primarily reflect the mechanical outcomes of movement rather than the underlying neuromuscular processes. Recent reviews of lower-limb exoskeleton sensing distinguish these mechanical (posterior) signals from earlier physiological signals, emphasizing that motion and foot–ground contact reflect the mechanical outcome of movement, whereas biological signals such as EMG are more directly linked to the neuromuscular processes underlying movement intention and execution [2]. Consequently, electromyography (EMG) has gained increasing attention as a complementary modality for gait-state estimation, with studies explicitly leveraging its anticipatory (ahead-of-motion) characteristics for gait phase recognition and kinematics prediction [8,23].

Building on this physiological advantage, both conventional machine learning and deep learning approaches have been extensively investigated for EMG-based gait recognition. Recent reviews of EMG classification indicate that handcrafted feature pipelines remain widely used, often combined with conventional supervised classifiers such as support vector machines, k-nearest neighbors, and random forests, particularly when dataset sizes are limited [18]. In gait-specific studies, Park et al. [24] systematically compared multiple classifiers, including decision trees, k-nearest neighbors, support vector machines, and neural networks, while Morbidoni et al. [25] applied deep learning for stance–swing classification during level-ground walking. Ling et al. [11] proposed a CNN-based domain-adaptive model to address gait phase recognition under varying walking speeds, Cai et al. [3] introduced an LDA-PSO-LSTM framework for seven-phase recognition, and Guerra et al. [8] employed a bidirectional LSTM model for stance–swing identification during exoskeleton-assisted walking. Collectively, these studies demonstrate a transition toward more adaptive temporal and deep learning architectures, while highlighting that conventional methods remain competitive and widely used in practical applications.

Despite these algorithmic advances, feature representation remains a critical and influential design factor. Recent reviews consistently categorize EMG features into time-domain (TD), frequency-domain (FD), and time–frequency (TF) groups [26,27]. Within this framework, TD features are widely used in gait phase recognition studies due to their ability to capture activation magnitude and timing, whereas FD features primarily characterize spectral properties associated with broader physiological states such as muscle fatigue. TF approaches aim to capture both temporal and spectral information but are less commonly adopted due to increased computational complexity and reduced interpretability [28,29]. This classification is directly relevant to gait phase recognition, where common TD descriptors—such as root mean square, mean absolute value, waveform length, zero-crossings, and slope-sign changes—remain widely used due to their simplicity and effectiveness. For example, Nazmi et al. [30] employed a TD feature set including root mean square, standard deviation, mean absolute value, integrated EMG, and waveform length, showing the continued effectiveness of established TD descriptors in recent gait phase classification studies.

Evidence from gait studies also suggests that task formulation significantly influences the effectiveness of feature-domain representations. In Nazmi et al. [30], binary stance–swing classification achieved an accuracy of approximately 0.98 across all features, suggesting that two-class phase discrimination can be effectively achieved using relatively simple feature representations. In contrast, Cai et al. addressed a more complex seven-phase recognition task using feature-combination selection across TD, FD, and TF domains, while Morbidoni et al. and Guerra et al. utilized deep learning architectures to model stance–swing dynamics directly from EMG signals [3,8,25]. Collectively, these findings indicate that the suitability of feature-domain representations is task-dependent, with simpler TD-based features often sufficient for binary classification, while more complex tasks may require richer or learned representations.

A second major issue is generalization across subjects, which remains a key challenge in EMG-based gait analysis. Di Nardo et al. demonstrated that intra-subject models outperform inter-subject models for EMG-based gait-event prediction, highlighting the significant impact of inter-individual variability [31]. A subsequent study by Di Nardo et al. further showed that both EMG signal processing strategies and sensor configuration complexity influence classification and prediction performance [32]. More recent work has attempted to address this limitation through domain adaptation and rapid recalibration techniques, particularly to accommodate variations such as walking speed and environmental conditions [11]. At a broader methodological level, recent research in wearable robotics emphasizes participant-based data splitting or leave-one-subject-out (LOSO) validation as more realistic evaluation strategies compared to random mixed-subject splits, as these approaches better assess performance on unseen users and reduce optimistic bias caused by cross-subject data leakage [33].

Overall, the literature supports EMG as a robust and promising modality for gait phase recognition. However, existing studies have not systematically compared time-domain and frequency-domain handcrafted features under controlled experimental conditions, particularly when classifier choice and subject-independent evaluation protocols are held constant. This limitation prevents definitive conclusions regarding the most suitable feature-domain representation for reliable stance–swing classification in practical deployment scenarios, thereby defining a focused and well-motivated research gap.

III. METHODS

A. Subject

Ten healthy male adults aged between 18 and 40 years participated in this study. This age range was selected to represent individuals with fully developed and stable neuromuscular control [34]. All participants reported no history of neurological disorders, musculoskeletal injuries, or gait impairments, ensuring that the recorded muscle activity reflected normal walking patterns. Prior to participation, all subjects provided written informed consent. The study protocol was reviewed and approved by the Universiti Teknologi Malaysia Research Ethics Committee (Approval No. UTMREC-2025-122).

B. Experimental Protocol

Each participant performed level-ground walking trials in a controlled laboratory environment. During each trial, participants walked along a straight, unobstructed pathway from a marked starting point to an end point and then returned to the starting location, as shown in Fig. 1. To minimize variability caused by footwear differences, all participants wore standardized shoes during the experiment. Participants were instructed to walk at their self-selected comfortable speed in order to preserve natural gait mechanics. This approach allows the recorded muscle activation patterns to represent natural locomotion behavior rather than constrained walking conditions.

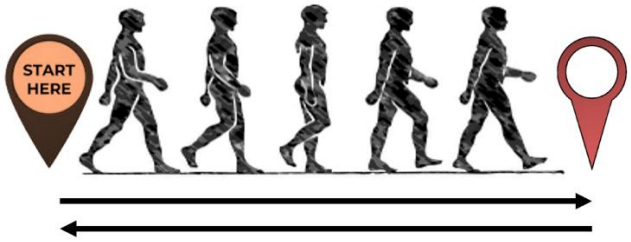


Fig. 1. Experimental protocol for walking trial.

C. Data Collection

EMG signals were recorded from two lower-limb muscles: the tibialis anterior (TA) and the medial gastrocnemius (MGAS). These muscles were selected because they play complementary roles in ankle control during walking. The TA primarily contributes to ankle dorsiflexion and is typically active during the swing phase to ensure foot clearance, whereas the MGAS contributes to plantarflexion and propulsion during the stance phase [35,36]. EMG electrodes were placed on this muscle based on the recommendations of Surface Electromyography for Non-Invasive Assessment (SENIAM) [37]. The electrodes were connected to a Biopac MP150 data acquisition system, and EMG signals were sampled at 1000 Hz, providing sufficient temporal resolution to capture rapid muscle activation changes during gait.

D. Gait Phase Labeling

The gait cycle was divided into two primary phases, which are stance and swing. The stance phase corresponds to the period during which the foot is in contact with the ground, while the swing phase refers to the period when the foot is lifted and moves forward to prepare for the next step [38]. Ground-truth labels for the gait phases were obtained through manual annotation based on visual inspection of the recorded walking sequences. A binary labeling scheme was adopted in which stance was assigned the label 1 and swing was assigned the label 0. In typical human walking, the stance phase occupies approximately 60% of the gait cycle, while the swing phase accounts for the remaining 40% [38]. This temporal distribution reflects the functional roles of the two phases, where stance provides support and propulsion while swing advances the limb for the subsequent step. Fig. 2 shows an example of the EMG signals recorded from the TA and MGAS muscles together with the corresponding stance–swing labels. These labeled gait phases were subsequently used as target classes for supervised classification after preprocessing and feature extraction.

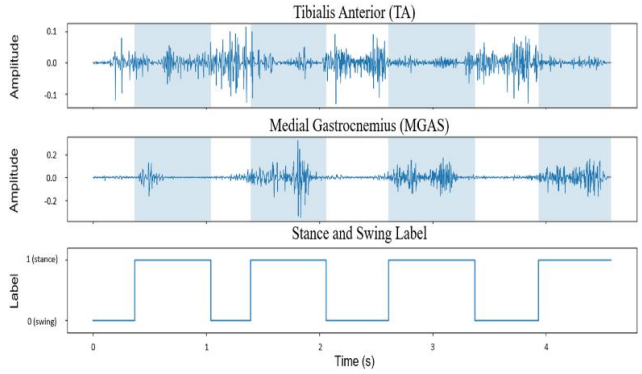


Fig. 2. Stance and swing labeling.

E. EMG Signal Preprocessing

The EMG signals were preprocessed using a sixth-order Butterworth band-pass filter with cutoff frequencies of 20–450 Hz to reduce noise and artifacts [39,40]. This frequency range preserves the dominant EMG spectral components while attenuating motion artifacts and high-frequency noise. After filtering, the EMG signals were segmented using a sliding window approach. Each analysis window had a duration of 200 ms with 50% overlap between consecutive windows, a configuration commonly adopted in EMG analysis to balance temporal resolution with statistical stability [41]. This corresponds to a step size of 100 ms between adjacent windows. For each analysis window, the corresponding gait phase label was determined using a majority voting strategy based on the stance–swing annotations within that window.

F. Feature Extraction

Two categories of features were investigated in this study: time-domain (TD) features and frequency-domain (FD) features. Feature extraction was performed independently for the TA and MGAS signals within each analysis window. The features extracted from both muscles were concatenated to form the final feature vector used for classification.

1) Time-domain features: Time-domain features characterize the amplitude and temporal properties of EMG signals. Six well-established TD features were employed, which are Mean Absolute Value (MAV), Standard Deviation (SD), Root Mean Square (RMS), Integrated EMG (iemg), Waveform Length (WL), and Zero Crossing (ZC), as Shown in Table I. These features have been widely validated in the

literature and are known to achieve high classification accuracy [42,43].

TABLE I. TIME DOMAIN (TD) FEATURE EXTRACTION

Time Domain (TD) features	For each window containing N signal samples *xi*, the following features were extracted.
Mean Absolute Value (MAV)	$MAV = \frac{1}{N}\sum_{i=1}^{N} \lvert x_i \rvert$
Standard Deviation (SD)	$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$, μ=mean of EMG signal
Root Mean Square (RMS)	$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$
Integrated EMG (IEMG)	$IEMG = \sum_{i=1}^{N} \lvert x_i \rvert$
Waveform Length (WL)	$WL = \sum_{i=1}^{N-1} \lvert x_{i+1} - x_i \rvert$
Zero Crossing (ZC)	$ZC = \sum_{n=1}^{N-1} \lvert sgn(x_i) - sgn(x_i - 1) \rvert$ where sgn(), $sgn(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$

2) Frequency-Domain features: Frequency-domain (FD) features describe the spectral properties of EMG signals. In this study, five widely used FD features were extracted, namely Mean Frequency (MNF), Median Frequency (MDF), Peak Frequency (PF), Total Spectral Power (TSP), and Spectral Entropy, as summarized in Table II. These features have also been widely validated in the literature on EMG-based classification tasks [44,45].

TABLE II. FREQUENCY DOMAIN (FD) FEATURE EXTRACTION

Frequency Domain (FD) features	The power spectral density (PSD) of each window was estimated using Welch's method. Let P(j) denote the power spectral density at the j-th frequency bin, where M is the total number of frequency bins.
Mean Frequency (MNF)	$MNF = \frac{\sum_{j=1}^{MDF} f_j P(j)}{\sum P(j)}$
Median Frequency (MDF)	$\sum_{j=1}^{MDF} P(j) = \sum_{j=MDF}^{M} P(j) = \frac{1}{2}\sum_{j=1}^{M} P(j)$
Peak Frequency (PF)	$PF = \max P(j), \quad j = 1,...,M$
Total Spectral Power (TTP)	$TTP = \sum_{j=1}^{M} P(j)$
Spectral Entropy (SpecEn)	$\text{SpecEn} = -\frac{\sum_{j=1}^{M} P(j)\log_2 P(j)}{\log_2 M}$

G. Support Vector Machine Classification

Support vector machine (SVM) is a supervised learning models that find an optimal hyperplane that separates the data into different classes by transforming it into a higher-dimensional space [46]. This transformation allows SVM to handle nonlinear decision boundaries by implicitly mapping inputs into high-dimensional feature spaces. In addition, SVMs often yield high accuracy and generalize well on high-dimensional data, making them popular for many classification problems [47,48]. In this study, gait phase classification was performed using an SVM classifier with a radial basis function (RBF) kernel. Before classification, the extracted feature vectors were standardized using z-score normalization so that all features contributed equally to the learning process. The classification pipeline, therefore, consisted of a feature scaling stage followed by SVM classification.

The SVM hyperparameters were configured as follows:

- Kernel: radial basis function (RBF)
- Kernel parameter: γ=0.005

The selection of the kernel parameter γ = 0.005 was based on preliminary empirical evaluation to balance model generalization and decision boundary smoothness. In the context of an RBF kernel, smaller γ values produce smoother decision boundaries and reduce the risk of overfitting, whereas larger values lead to more complex boundaries that may degrade generalization performance, particularly under subject-independent evaluation [49].

H. Model Evaluation

To evaluate the generalization capability of the classifier across participants, a leave-one-subject-out (LOSO) cross-validation strategy was employed. In this approach, the data from one subject was used as the test set while the remaining subjects were used to train the classifier. The process was repeated until each subject had served once as the test subject. Classification performance was assessed using accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC), as shown in Table III. The F1-score represents the harmonic mean of precision and recall, while the macro-averaged F1-score (Macro-F1) was computed as the average of class-wise F1-scores to ensure balanced evaluation across stance and swing classes. MCC was utilized as a metric to capture agreement beyond chance while providing a balanced evaluation of classification performance across both classes.

TABLE III. PERFORMANCE METRIC

Performance Metric	Formula
Accuracy	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\text{Precision} = \frac{TP}{TP + FP}$
Recall	$\text{Recall} = \frac{TP}{TP + FN}$
F1-score	$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

Macro F1	$\text{Macro F1} = \frac{F1_{\text{swing}} + F1_{\text{stance}}}{2}$
MCC	$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

IV. RESULTS

A. Overall Classification Performance (Time vs. Frequency Domain Features)

The overall classification performance of the time-domain (TD) and frequency-domain (FD) feature sets evaluated using leave-one-subject-out (LOSO) cross-validation is summarized in Table IV and illustrated in Fig. 3. The TD feature set achieved a higher classification accuracy of 0.813 ± 0.112, outperforming the FD feature set with an accuracy of 0.712 ± 0.077. A similar trend was observed for the macro-averaged F1-score (Macro-F1), where TD features achieved 0.812 ± 0.114, compared to 0.708 ± 0.079 for FD features. Furthermore, the Matthews correlation coefficient (MCC) also indicated superior performance for TD features (0.672 ± 0.178) relative to FD features (0.448 ± 0.159).

TABLE IV. OVERALL PERFORMANCE COMPARISON OF TIME-DOMAIN (TD) AND FREQUENCY-DOMAIN (FD) FEATURE SETS USING LOSO CROSS-VALIDATION BASED ON FIG. 3.

Feature Mode	Accuracy	Macro-F1	MCC
TD	0.813 ± 0.112	0.812 ± 0.114	0.672 ± 0.178
FD	0.712 ± 0.077	0.708 ± 0.079	0.448 ± 0.159

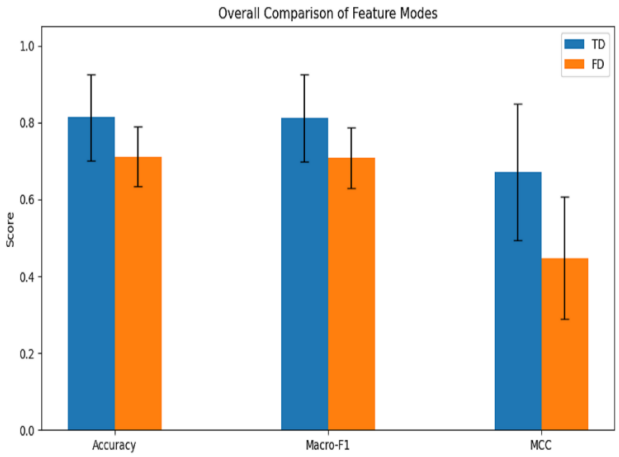


Fig. 3. Comparison of classification performance between time-domain (TD) and frequency-domain (FD) EMG feature sets using SVM under leave-one-subject-out cross-validation. Error bars represent standard deviation across subjects.

B. Per-Class Performance and Confusion Matrix Analysis

To further examine class-specific behavior, Fig. 4 presents a heatmap of precision, recall, and F1-score for both swing and stance phases. For the TD feature set, swing phase detection achieved a recall of 0.935 and an F1-score of 0.822, whereas stance detection exhibited a recall of 0.729 and a precision of 0.929. For the FD feature set, both phases exhibited lower performance (swing F1: 0.687, stance F1: 0.729), indicating reduced classification effectiveness.

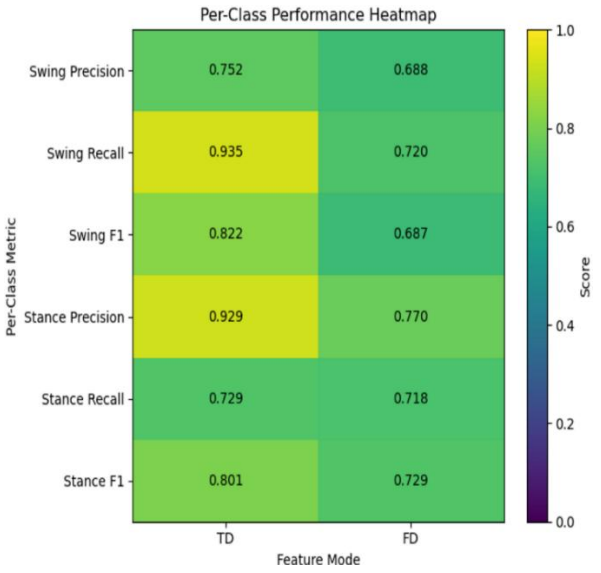


Fig. 4. Per-class performance heatmap.

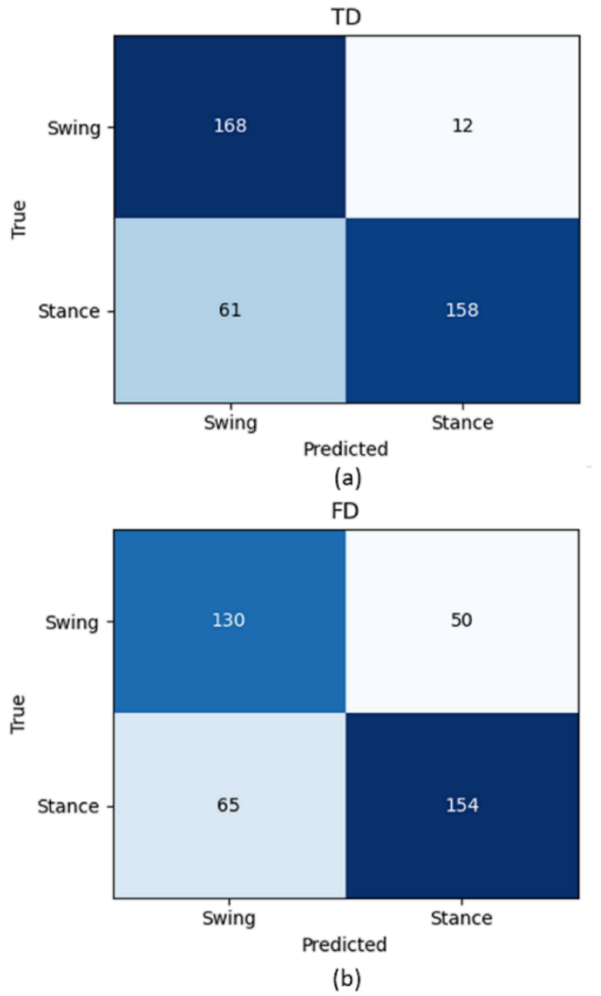


Fig. 5. Confusion matrices across feature modes: (a) Time-domain (TD) (b) Frequency-domain (FD).

Fig. 5 presents the confusion matrices, for the (a) time-domain (TD) and (b) frequency-domain (FD) feature sets,

providing further insight into classification errors. Most predictions are concentrated along the diagonal, indicating correct classification of both gait phases. However, a noticeable proportion of stance samples are misclassified as swing, consistent with the lower stance recall observed in Fig. 4. Importantly, Fig. 5(a) exhibits fewer off-diagonal errors compared to Fig. 5(b), confirming its superior classification performance. This reduction in misclassification highlights the effectiveness of TD features in capturing phase-specific amplitude variations during gait transitions.

V. Discussion

A. Interpretation of Feature Performance

In Table IV, the results demonstrate that TD features provide stronger discriminative information for gait phase classification compared to FD features. This is likely because time-domain descriptors capture transient neuromuscular activation patterns, whereas frequency-domain features primarily represent spectral distributions and may be less sensitive to temporal variations. This observation aligns with previous studies reporting that TD features often outperform FD features in dynamic EMG classification tasks [50,51]. The higher MCC values further indicate that TD features maintain more balanced classification performance across classes.

B. Biomechanical Interpretation of Class Differences

The higher recall for the swing phase observed in Fig. 4 can be explained by gait biomechanics. During swing, the tibialis anterior (TA) exhibits a distinct high-amplitude activation burst associated with ankle dorsiflexion and foot clearance [52]. This produces clear and temporally localized EMG patterns, making the phase easier to detect. In contrast, the stance phase involves more distributed and overlapping muscle activations, including contributions from the gastrocnemius and soleus. This results in increased signal variability and reduced separability. Similar findings have been reported in prior work, where swing phases are often classified more accurately than stance phases despite their shorter duration [25].

C. Implications for EMG-Based Gait Phase Detection

The findings highlight the critical role of feature representation in EMG-based gait phase classification. Across all evaluation metrics, TD features consistently outperformed FD features, demonstrating their effectiveness in capturing neuromuscular activation dynamics during locomotion. This result is physiologically justified, as EMG signals predominantly reflect amplitude modulation of muscle activity, which is more directly represented in the time domain [53]. From an application perspective, TD features provide a computationally efficient and reliable basis for real-time gait phase detection. Their strong classification performance indicates that they can effectively support control strategies in assistive technologies, such as powered prostheses and lower-limb exoskeletons, where accurate phase detection is required to synchronize mechanical assistance with user movement [24].

A key limitation of the present framework is that each analysis window is treated as an independent sample, without explicitly modeling temporal dependencies between consecutive gait phases. Consequently, the model does not enforce physiologically consistent phase transitions, which may lead to misclassification near transition boundaries. Recent studies have demonstrated that sequence-aware models, such as recurrent neural networks and temporal convolutional networks, can improve gait phase recognition by capturing temporal continuity and transition constraints [8,54]. Another limitation of the present study is the use of only two lower-limb muscles, namely the tibialis anterior and medial gastrocnemius. Although these muscles provide strong discriminative information due to their roles in dorsiflexion and plantarflexion, gait is inherently governed by coordinated activation across multiple muscle groups. As both muscles are located at the distal segment of the lower limb, the current framework primarily reflects ankle-level control and does not capture the contributions of proximal muscles involved in hip and knee stabilization. Consequently, this limited muscle representation may restrict the generalizability of the findings to more complex locomotion scenarios that require multi-joint coordination and whole-body control. Previous studies have shown that incorporating additional muscles or muscle synergy information can improve classification robustness and better capture the complexity of locomotion [24,55].

Overall, the results demonstrate that TD features provide a physiologically meaningful and computationally efficient representation for gait phase detection. These characteristics make them well-suited for implementation in real-time wearable systems, supporting both assistive device control and rehabilitation-oriented gait monitoring applications.

VI. Conclusion

This study investigated the effectiveness of time-domain (TD) and frequency-domain (FD) electromyography (EMG) features for stance–swing gait phase classification using signals from the tibialis anterior (TA) and medial gastrocnemius (MGAS) muscles during walking. A support vector machine classifier was evaluated under a subject-independent leave-one-subject-out framework to assess generalization performance. The results consistently demonstrate that TD features outperform FD features across all evaluation metrics, including classification accuracy, macro-averaged F1-score, and Matthews correlation coefficient. These findings indicate that amplitude-based temporal characteristics of EMG signals provide more discriminative information for stance–swing classification than spectral representations under the evaluated conditions. Despite these findings, several limitations should be acknowledged. First, the study was conducted on a relatively small dataset consisting of ten healthy participants, which may limit the statistical generalizability of the results. Second, the observed variability in classification performance, particularly in MCC values, reflects inter-subject differences in EMG signal characteristics, including muscle activation patterns and signal quality. Third, the analysis was restricted to two muscles (TA and MGAS) and to predefined TD and FD feature sets, which may not fully capture the complexity of multi-muscle coordination during gait. From an application perspective, TD features offer a computationally efficient and physiologically meaningful representation for real-time gait phase detection. However, the findings should be interpreted within the scope of the experimental setup and feature design choices, particularly under controlled laboratory conditions. These constraints

highlight the need for further validation before deployment in real-world assistive systems.

Future work should focus on improving the robustness, scalability, and real-world applicability of EMG-based gait phase classification systems. One important direction is the inclusion of larger and more diverse participant populations, including individuals with gait impairments, to better evaluate generalization across varying neuromuscular conditions. In addition, future studies should investigate multi-muscle configurations to capture a more comprehensive representation of lower-limb coordination. Another key direction involves the integration of temporal modeling approaches. The current study treats EMG windows independently and does not incorporate sequential dependencies between gait phases. Incorporating sequence-aware models, such as recurrent neural networks or temporal convolutional architectures, may improve classification stability and reduce misclassification during phase transitions. Furthermore, while this study compared TD and FD features independently, combining both feature domains may provide complementary information that enhances classification performance. Hybrid feature representations that integrate temporal and spectral characteristics could offer improved robustness, particularly under noisy or variable conditions. Finally, future research should evaluate the proposed approach under real-world conditions, including varying walking speeds, terrains, and sensor noise levels, to assess its suitability for deployment in wearable assistive technologies and continuous gait monitoring applications.

ACKNOWLEDGMENT

The authors would like to acknowledge the contributions of all participants involved in this study. This work was supported by the Universiti Teknologi Malaysia, Geran Penyelidikan Kecerdasan Buatan, 44H54.

REFERENCES

[1] ADDIN Mendeley Bibliography CSL_BIBLIOGRAPHY R. Soldi, B.M.V. Guerra, S. Sozzi, L. Russo, S. Pizzocaro, R. Baptista, A.M. De Nunzio, M. Schmid, S. Ramat, Deep Learning Models Optimization for Gait Phase Identification from EMG Data During Exoskeleton-Assisted Walking, Biomimetics 10 (2025) 1–25, https://doi.org/10.3390/biomimetics10090617.

[2] Y. Sun, Y. Tang, J. Zheng, D. Dong, X. Chen, L. Bai, From sensing to control of lower limb exoskeleton: a systematic review, Annu. Rev. Control 53 (2022) 83–96, https://doi.org/10.1016/j.arcontrol.2022.04.003.

[3] S. Cai, D. Chen, B. Fan, M. Du, G. Bao, G. Li, Gait phases recognition based on lower limb sEMG signals using LDA-PSO-LSTM algorithm, Biomed. Signal Process. Control 80 (2023) 104272, https://doi.org/10.1016/j.bspc.2022.104272.

[4] O. Coser, C. Tamantini, P. Soda, L. Zollo, AI-based methodologies for exoskeleton-assisted rehabilitation of the lower limb: a review, Front. Robot. AI 11 (2024) 1–19, https://doi.org/10.3389/frobt.2024.1341580.

[5] J. de Miguel-Fernández, J. Lobo-Prat, E. Prinsen, J.M. Font-Llagunes, L. Marchal-Crespo, Control strategies used in lower limb exoskeletons for gait rehabilitation after brain injury: a systematic review and analysis of clinical effectiveness, J. Neuroeng. Rehabil. 20 (2023) 1–28, https://doi.org/10.1186/s12984-023-01144-5.

[6] H. Prasanth, M. Caban, U. Keller, G. Courtine, A. Ijspeert, H. Vallery, J. von Zitzewitz, Wearable sensor-based real-time gait detection: A systematic review, Sensors 21 (2021) 2727, https://doi.org/10.3390/s21082727.

[7] Z. Zhang, Z. Wang, H. Lei, W. Gu, Gait phase recognition of lower limb exoskeleton system based on the integrated network model, Biomed. Signal Process. Control 76 (2022) 103693, https://doi.org/10.1016/j.bspc.2022.103693.

[8] B.M.V. Guerra, M. Schmid, S. Sozzi, S. Pizzocaro, A.M. De Nunzio, S. Ramat, A Recurrent Deep Network for Gait Phase Identification from EMG Signals During Exoskeleton-Assisted Walking, Sensors 24 (2024) 6666, https://doi.org/10.3390/s24206666.

[9] J. Liu, C. Wang, B. He, P. Li, X. Wu, Metric Learning for Robust Gait Phase Recognition for a Lower Limb Exoskeleton Robot Based on sEMG, IEEE Trans. Med. Robot. Bionics 4 (2022) 472–479, https://doi.org/10.1109/TMRB.2022.3166543.

[10] X. Shi, X. Zhang, P. Qin, L. Huang, Y. Zhu, Z. Yang, Gait Phase Recognition in Multi-Task Scenarios Based on sEMG Signals, Biosensors 15 (2025) 1–21, https://doi.org/10.3390/bios15050305.

[11] Z.Q. Ling, J.C. Chen, G.Z. Cao, Y.P. Zhang, L.L. Li, W.X. Xu, S. Bin Cao, A Domain Adaptive Convolutional Neural Network for sEMG-Based Gait Phase Recognition Against to Speed Changes, IEEE Sens. J. 23 (2023) 2565–2576, https://doi.org/10.1109/JSEN.2022.3228320.

[12] P. Zhang, J. Zhang, A. Elsabbagh, Lower Limb Motion Intention Recognition Based on sEMG Fusion Features, IEEE Sens. J. 22 (2022) 7005–7014, https://doi.org/10.1109/JSEN.2022.3146446.

[13] Q. Liu, W. Sun, N. Peng, W. Meng, S.Q. Xie, DCNN-SVM-Based Gait Phase Recognition With Inertia, EMG, and Insole Plantar Pressure Sensing, IEEE Sens. J. 24 (2024) 28869–28878, https://doi.org/10.1109/JSEN.2024.3435884.

[14] S.T. Kwak, Y.H. Chang, Fascicle dynamics of the tibialis anterior muscle reflect whole-body walking economy, Sci. Rep. 13 (2023) 1–11, https://doi.org/10.1038/s41598-023-31501-2.

[15] S.H. Liu, A.K. Sharma, B.Y. Wu, X. Zhu, C.J. Chang, J.J. Wang, Estimating gait parameters from sEMG signals using machine learning techniques under different power capacity of muscle, Sci. Rep. 15 (2025) 1–15, https://doi.org/10.1038/s41598-025-95973-0.

[16] E. Anselmino, A. Mazzoni, S. Micera, EMG-based prediction of step direction for a better control of lower limb wearable devices, Comput. Methods Programs Biomed. 254 (2024) 108305, https://doi.org/10.1016/j.cmpb.2024.108305.

[17] J.H. Sul, L. Piyathilaka, D. Moratuwage, S. Dunu Arachchige, A. Jayawardena, G. Kahandawa, D.M.G. Preethichandra, Electromyography Signal Acquisition, Filtering, and Data Analysis for Exoskeleton Development, Sensors 25 (2025) 1–31, https://doi.org/10.3390/s25134004.

[18] S.M. Sid'El Moctar, I. Rida, S. Boudaoud, Comprehensive Review of Feature Extraction Techniques for sEMG Signal Classification: From Handcrafted Features to Deep Learning Approaches, Irbm 45 (2024) 100866, https://doi.org/10.1016/j.irbm.2024.100866.

[19] L. Quesada, D. Verdel, O. Bruneau, B. Berret, M.A. Amorim, N. Vignais, EMG feature extraction and muscle selection for continuous upper limb movement regression, Biomed. Signal Process. Control 103 (2025) 107323, https://doi.org/10.1016/j.bspc.2024.107323.

[20] P.N. Gunaratne, H. Tamura, An EMG-Based GRU Model for Estimating Foot Pressure to Support Active Ankle Orthosis Development, Sensors 25 (2025) 1–22, https://doi.org/10.3390/s25113558.

[21] H.T.T. Vu, D. Dong, H.L. Cao, T. Verstraten, D. Lefeber, B. Vanderborght, J. Geeroms, A review of gait phase detection algorithms for lower limb prostheses, Sensors (Switzerland) 20 (2020) 1–19, https://doi.org/10.3390/s20143972.

[22] V.M. Santos, B.B. Gomes, M.A. Neto, A.M. Amaro, A Systematic Review of Insole Sensor Technology: Recent Studies and Future Directions, Appl. Sci. 14 (2024), https://doi.org/10.3390/app14146085.

[23] B. Wei, Z. Ding, C. Yi, H. Guo, Z. Wang, J. Zhu, F. Jiang, A Novel sEMG-Based Gait Phase-Kinematics-Coupled Predictor and Its Interaction With Exoskeletons, Front. Neurorobot. 15 (2021) 1–14, https://doi.org/10.3389/fnbot.2021.704226.

[24] H. Park, S. Han, J. Sung, S. Hwang, I. Youn, S.J. Kim, Classification of gait phases based on a machine learning approach using muscle synergy, Front. Hum. Neurosci. 17 (2023) 1–12, https://doi.org/10.3389/fnhum.2023.1201935.

[25] C. Morbidoni, A. Cucchiarelli, S. Fioretti, F. Di Nardo, A deep learning approach to EMG-based classification of gait phases during level ground walking, Electron. 8 (2019), https://doi.org/10.3390/electronics8080894.

[26] J. Sun, G. Liu, Y. Sun, K. Lin, Z. Zhou, J. Cai, Application of Surface Electromyography in Exercise Fatigue: A Review, Front. Syst. Neurosci. 16 (2022) 1–8, https://doi.org/10.3389/fnsys.2022.893275.

[27] L. Piyathilaka, J.H. Sul, S. Dunu Arachchige, A. Jayawardena, D. Moratuwage, Advances in EMG Signal Processing and Pattern Recognition: Techniques, Challenges, and Emerging Applications, Electron. 15 (2026) 1–40, https://doi.org/10.3390/electronics15030590.

[28] S. Muceli, R. Merletti, Tutorial. Frequency analysis of the surface EMG signal: Best practices, J. Electromyogr. Kinesiol. 79 (2024) 102937, https://doi.org/10.1016/j.jelekin.2024.102937.

[29] T. Song, Z. Yan, S. Guo, Y. Li, X. Li, F. Xi, Review of sEMG for Robot Control: Techniques and Applications, Appl. Sci. 13 (2023), https://doi.org/10.3390/app13179546.

[30] N. Nazmi, S.I. Yamamoto, M.A.S. Rohim, E.F. Shair, Classification of Gait Phases by Using SVM and ANN Based on EMG Signals, 2024 IEEE Symp. Ind. Electron. Appl. ISIEA 2024 (2024) 1–6, https://doi.org/10.1109/ISIEA61920.2024.10607348.

[31] F. Di Nardo, C. Morbidoni, G. Mascia, F. Verdini, S. Fioretti, Intra-subject approach for gait-event prediction by neural network interpretation of EMG signals, Biomed. Eng. Online 19 (2020) 1–20, https://doi.org/10.1186/s12938-020-00803-1.

[32] F. Di Nardo, C. Morbidoni, A. Cucchiarelli, S. Fioretti, Influence of EMG-signal processing and experimental set-up on prediction of gait events by neural network, Biomed. Signal Process. Control 63 (2021) 102232, https://doi.org/10.1016/j.bspc.2020.102232.

[33] O. Coser, C. Tamantini, M. Tortora, L. Furia, R. Sicilia, L. Zollo, P. Soda, Deep learning for human locomotion analysis in lower-limb exoskeletons: a comparative study, Front. Comput. Sci. 7 (2025), https://doi.org/10.3389/fcomp.2025.1597143.

[34] T. Hortobágyi, T. Vetrovsky, J.S. Brach, M. van Haren, K. Volesky, R. Radaelli, P. Lopez, U. Granacher, Effects of Exercise Training on Muscle Quality in Older Individuals: A Systematic Scoping Review with Meta-Analyses, Sport. Med. - Open 9 (2023), https://doi.org/10.1186/s40798-023-00585-5.

[35] S. Bajelan, W.A. Sparrow, R. Begg, The ankle dorsiflexion kinetics demand to increase swing phase foot-ground clearance: implications for assistive device design and energy demands, J. Neuroeng. Rehabil. 21 (2024) 1–12, https://doi.org/10.1186/s12984-024-01394-x.

[36] B. Kibushi, Muscle coordination patterns in regulation of medial gastrocnemius activation during walking, Hum. Mov. Sci. 90 (2023) 103116, https://doi.org/10.1016/j.humov.2023.103116.

[37] H.J. Hermens, B. Freriks, R. Merletti, D. Stegeman, J. Blok, G. Rau, C. Disselhorst-Klug, G. Hägg, European Recommendations for Surface ElectroMyoGraphy Results of the SENIAM project, Roessingh Research and Development, 1999. https://books.google.com.my/books?id=w7HgOwAACAAJ.

[38] P. Natarajan, R.D. Fonseka, S. Kim, C. Betteridge, M. Maharaj, R.J. Mobbs, Analysing gait patterns in degenerative lumbar spine diseases: a literature review, J. Spine Surg. 8 (2022) 139–148, https://doi.org/10.21037/jss-21-91.

[39] F. Zhang, H. Huang, Source selection for real-time user intent recognition toward volitional control of artificial legs, IEEE J. Biomed. Heal. Informatics 17 (2013) 907–914, https://doi.org/10.1109/JBHI.2012.2236563.

[40] L. Chen, Y. Wang, X. Lu, J. Geng, T. Feng, Modelling and analysis of myoelectric networks based on lower limb motor synergies, IEEE Trans. Neural Syst. Rehabil. Eng. 33 (2025) 3910–3921, https://doi.org/10.1109/TNSRE.2025.3613408.

[41] H. Wang, Q. Tao, X. Zhang, Ensemble Learning Method for the Continuous Decoding of Hand Joint Angles, Sensors 24 (2024) 1–17, https://doi.org/10.3390/s24020660.

[42] C. Angelidou, P. Artemiadis, On Predicting Transitions to Compliant Surfaces in Human Gait via Neural and Kinematic Signals, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 2214–2223, https://doi.org/10.1109/TNSRE.2023.3272355.

[43] K.H. Lee, J.Y. Min, S. Byun, Electromyogram-based classification of hand and finger gestures using artificial neural networks, Sensors 22 (2022), https://doi.org/10.3390/s22010225.

[44] A. Phinyomark, A. Nuidod, P. Phukpattaranont, C. Limsakul, Feature extraction and reduction of wavelet transform coefficients for EMG pattern classification, Elektron. Ir Elektrotechnika 122 (2012) 27–32, https://doi.org/10.5755/j01.eee.122.6.1816.

[45] P. T., V.K. Elumalai, B. E., Hand gesture classification framework leveraging the entropy features from sEMG signals and VMD augmented multi-class SVM, Expert Syst. Appl. 238 (2024) 121972, https://doi.org/10.1016/j.eswa.2023.121972.

[46] A.C. Tsai, J.J. Luh, T. Te Lin, A novel STFT-ranking feature of multi-channel EMG for motion pattern recognition, Expert Syst. Appl. 42 (2015) 3327–3341, https://doi.org/10.1016/j.eswa.2014.11.044.

[47] R. Guido, S. Ferrisi, D. Lofaro, D. Conforti, An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review, Inf. 15 (2024), https://doi.org/10.3390/info15040235.

[48] P. Thanh Noi, M. Kappas, Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery, Sensors (Basel). 18 (2017), https://doi.org/10.3390/s18010018.

[49] A. Mastropietro, C. Feldmann, J. Bajorath, Calculation of exact Shapley values for explaining support vector machine models using the radial basis function kernel, Sci. Rep. 13 (2023) 1–11, https://doi.org/10.1038/s41598-023-46930-2.

[50] M. Faisal, I. Khosa, A. Waris, S.O. Gilani, M.J. Khan, F. Hazzazi, M.A. Ijaz, Optimizing the impact of time domain segmentation techniques on upper limb EMG decoding using multimodal features, PLoS One 20 (2025) 1–21, https://doi.org/10.1371/journal.pone.0322580.

[51] S. Mendez-Moreno, L. Espinosa, O. Vital-Ochoa, R. Espinosa-Tanguma, J. Acosta-Elias, Segmentation of Surface Electromyography Signals: A Comparative Analysis of Time and Frequency Domain Methods, Comput. y Sist. 28 (2024) 1783–1797, https://doi.org/10.13053/CyS-28-4-4874.

[52] C.A. Byrne, D.T. O'Keeffe, A.E. Donnelly, G.M. Lyons, Effect of walking speed changes on tibialis anterior EMG during healthy gait for FES envelope design in drop foot correction, J. Electromyogr. Kinesiol. 17 (2007) 605–616, https://doi.org/10.1016/j.jelekin.2006.07.008.

[53] R.K. Mani, B. Nagaraj, Diagnosis and classification of neuromuscular disorders using Bi-LSTM optimized with grey Wolf optimizer for EMG signals, Sci. Rep. 15 (2025) 1–17, https://doi.org/10.1038/s41598-025-03766-2.

[54] T. Li, W. Ma, Y. Zheng, X. Fan, G. Yang, L. Wang, Z. Li, A survey on gait recognition against occlusion: taxonomy, dataset and methodology, PeerJ Comput. Sci. 10 (2024) 1–55, https://doi.org/10.7717/PEERJ-CS.2602.

[55] S. Wang, K. Hase, T. Funato, Computational prediction of muscle synergy using a finite element framework for a musculoskeletal model on lower limb, Front. Bioeng. Biotechnol. 11 (2023) 1–13, https://doi.org/10.3389/fbioe.2023.1130219.

A Framework for Digital Technology and AI Adoption in Slovak Firms: Evidence from Qualitative Analysis

Martina Chrancokova[1], Ludmila Mitkova[2], Lukas Vartiak[3]
Faculty of Management, Comenius University Bratislava, Bratislava, Slovakia[1]
Institute for Forecasting-Centre of Social and Psychological Sciences, Slovak Academy of Sciences, Bratislava, Slovakia[1]
Faculty of Management, Comenius University Bratislava, Bratislava, Slovakia[2]
Faculty of Social and Economic Sciences, Comenius University Bratislava, Bratislava, Slovakia[3]

***Abstract*—This study examines the adoption of digital technologies and artificial intelligence (AI) in Slovak firms, with particular attention to technological integration, employee adaptation, and organizational change. The study is based on qualitative semi-structured interviews conducted with managers and employees from 20 companies across multiple sectors and firm-size categories. The data were analyzed using thematic analysis. The findings identify three levels of digital adoption: basic digitization, process automation, and AI-supported adoption. Basic digital tools were reported across all firms, while advanced forms of adoption were concentrated mainly in IT and manufacturing companies. AI use remained limited and was typically confined to exploratory applications such as chatbots, automated support, or data-processing assistance. Across most cases, employees initially responded to digital change with hesitation or resistance, followed by gradual adaptation through practice-based and informal workplace learning. The results further indicate that digitalization was associated primarily with task reallocation, workflow optimization, and role redesign, whereas direct workforce reduction was reported only in isolated cases. Based on these findings, the study develops a conceptual framework linking technological adoption, employee adaptation, organizational restructuring, and sectoral context. Given the qualitative and exploratory nature of the research, the findings should be interpreted as analytically transferable rather than statistically generalizable. The study contributes to the literature by providing firm-level evidence from a Central and Eastern European context and by proposing a structured interpretation of digital and AI adoption as a multi-level organizational process.**

Keywords*—Digital technology adoption; artificial intelligence; digital transformation; process automation; thematic analysis; Slovak firms

I. Introduction

The rapid digitalization of economies and societies has fundamentally reshaped how organizations operate, compete, and create value. Digital technologies, including artificial intelligence (AI), are increasingly embedded in organizational processes, transforming operational models, decision-making practices, and workforce structures. As a result, firms that effectively adopt and integrate digital technologies gain competitive advantages, while those with limited adaptability risk declining productivity and reduced market relevance.

Despite the growing importance of digital transformation, existing research has primarily focused on technological and infrastructural aspects of digitalization, often emphasizing large organizations or technologically advanced economies. Less attention has been paid to how companies in Central and Eastern Europe (CEE), particularly small and medium-sized enterprises, interpret and implement digital transformation in practice. Moreover, prior studies frequently treat digital adoption as a purely technological process, overlooking the organizational and human dimensions, such as employee adaptation, managerial practices, and workplace learning, which are critical for successful implementation.

In the Slovak context, digital transformation has become a strategic priority, supported by national and European initiatives such as the Digital Economy and Society Index (DESI). However, empirical evidence on how Slovak companies perceive digitalization, adopt digital and AI technologies, and manage associated organizational changes remains limited. Existing studies tend to provide macro-level indicators of digital readiness but offer insufficient insight into firm-level processes, particularly regarding the interaction between technological adoption and employee adaptation.

Despite these developments, important gaps remain in the existing literature. First, empirical research on digital transformation in Central and Eastern Europe (CEE) remains limited at the firm level, particularly in relation to small and medium-sized enterprises. Second, prior studies tend to examine technological adoption and organizational or human factors separately, rather than as interconnected processes. Third, there is a lack of qualitative, process-oriented evidence capturing how firms transition between different stages of digital adoption in practice.

The aim of this study is to examine how Slovak companies adopt and implement digital technologies and artificial intelligence, with a particular focus on the interaction between technological adoption, employee adaptation, and organizational change. By addressing these dimensions simultaneously, the study seeks to provide a more integrated understanding of digital transformation as a multi-dimensional organizational process.

To achieve this aim, the study draws on qualitative semi-structured interviews conducted with managers and employees from 20 companies across diverse sectors.

The study is guided by the following research questions:

- How do Slovak companies perceive digitalization and digital transformation?
- How do companies adopt and use artificial intelligence and other digital tools?
- How do employees react to digital changes, and how do they adapt?
- What organizational effects does digitalization produce in Slovak companies?

This study contributes to the literature in three main ways. First, it provides empirical insights into digital and AI adoption in a CEE context, which remains underrepresented in current research. Second, it develops a structured understanding of digital adoption by identifying three levels of technological integration—basic digitization, process automation, and AI-supported adoption. Third, it proposes a conceptual framework that integrates technological, organizational, and human dimensions of digital transformation. Given the exploratory nature of the study and the qualitative design, the findings should be interpreted as analytically transferable rather than statistically generalizable.

II. Literature Review

Digitalization and digital transformation have become central concepts in contemporary organizational research [34]; however, their conceptual boundaries remain fluid and subject to ongoing debate. In the management and information systems literature, Digitalization is commonly understood as the use of digital technologies and data to improve existing processes and enhance operational efficiency [27], [10]. In contrast, digital transformation represents a broader, strategic reconfiguration of organizational structures, business models, and value creation processes [33]. While this distinction is widely acknowledged, empirical studies suggest that organizations often fail to differentiate clearly between these concepts in practice, frequently adopting a technology-driven rather than strategy-driven perspective. However, despite extensive conceptual development, the existing literature remains fragmented and often lacks empirical validation at the organizational level, particularly in less-studied regional contexts.

Recent literature increasingly emphasizes that digital transformation is not solely a technological process but a socio-technical phenomenon integrating technological innovation with organizational and human dimensions [32], [26]. In this context, digital transformation involves changes in leadership, organizational culture, and employee competencies, alongside the implementation of digital tools. However, despite this broader conceptualization, a significant portion of empirical research continues to prioritize technological adoption while underestimating the role of human capital and organizational learning processes. Nevertheless, many empirical studies continue to operationalize digital transformation primarily through technological indicators, thereby underestimating the complexity of human and organizational dynamics involved in the transformation process.

The importance of human capital in digital transformation has been widely documented. Studies show that employees' digital skills and adaptability are critical determinants of successful technology adoption and organizational innovation [29], [23]. Similarly, [16] argue that digital transformation is driven more by organizational strategy and leadership than by technology itself. Nevertheless, in practice, many firms—particularly small and medium-sized enterprises (SMEs)—struggle to align technological investments with workforce development, resulting in fragmented or incomplete transformation processes. This suggests that, particularly in SMEs, digital transformation is frequently implemented in a fragmented manner, with limited integration between technological investments and workforce development strategies.

Sectoral and regional differences further complicate the adoption of digital technologies. Research indicates that technologically intensive industries, such as manufacturing and information technology, tend to lead digital transformation due to greater access to financial resources, technological expertise, and innovation capabilities [11], [5]. In contrast, service-oriented sectors and smaller organizations often face resource constraints, skill shortages, and higher levels of uncertainty, which slow down the adoption of advanced technologies such as artificial intelligence. The likelihood of AI adoption is strongly influenced by firm-level characteristics, particularly organizational size and sectoral orientation. Large enterprises demonstrate significantly higher adoption rates compared to small and medium-sized enterprises (SMEs), primarily due to greater access to financial resources, technological expertise, and organizational capabilities [28]. Sectoral differences are also pronounced, with firms in the ICT sector exhibiting higher adoption rates, while industries such as agriculture, construction, and public services lag behind [12].

The literature identifies several key drivers of AI adoption. These include the need to address skill shortages in areas such as IT and data analytics, organizational growth and performance pressures, and participation in innovation ecosystems that facilitate access to knowledge and resources [12], [13]. At the same time, firms are primarily motivated by operational goals, such as process optimization, cost reduction, and service innovation.

However, the adoption process is constrained by a range of barriers. Absolute barriers include the lack of financial resources, insufficient technological infrastructure, and limited availability of skilled personnel [20]. In addition, relative barriers—such as data quality issues, integration challenges, and limited managerial support—affect firms already engaged in AI implementation. These findings highlight that AI adoption is not only a technological decision but also a complex organizational process shaped by internal capabilities and external constraints.

In the European context, and particularly in Central and Eastern Europe (CEE), digital transformation is shaped by structural and institutional factors, including disparities in digital skills, infrastructure, and innovation ecosystems. The Digital Economy and Society Index (DESI) highlights persistent gaps

in digital readiness across EU member states, with countries such as Slovakia exhibiting moderate performance but lagging in areas such as advanced digital skills and enterprise-level technology adoption [8]. While DESI provides valuable macro-level insights, it offers limited understanding of how firms interpret and operationalize digital transformation at the organizational level. Recent empirical evidence indicates that the adoption of artificial intelligence (AI) among European enterprises remains relatively limited, although it is gradually increasing. As of 2023, approximately 8% of enterprises in the European Union reported using at least one AI technology, representing a modest increase compared to 7.6% in 2021 [28]. However, significant disparities persist across countries. While Nordic economies such as Denmark and Sweden report adoption rates exceeding 35%, Central and Eastern European countries demonstrate substantially lower levels of implementation [13], [28].

In the Slovak context, AI adoption increased from 5.2% in 2021 to 7.0% in 2023, yet remains consistently below the EU average across all firm-size categories [28], [18]. These findings suggest that structural and institutional constraints continue to limit the diffusion of advanced digital technologies in the region.

Another important dimension of digital transformation concerns employee reactions to technological change. The literature consistently identifies resistance to change as a common response, particularly when new technologies disrupt established work routines [31], [24]. However, research also shows that such resistance is often temporary and can be mitigated through effective change management, communication, and training [19]. More recent studies highlight the role of informal, workplace-based learning as a mechanism through which employees gradually adapt to digital tools, especially in resource-constrained environments [7].

Despite these advances, several important limitations remain in the literature. First, empirical research on digital transformation in CEE countries remains relatively limited, particularly at the firm level. Second, existing studies often examine technological adoption and human adaptation as separate phenomena, rather than as interdependent processes shaping organizational outcomes. Third, most empirical research relies on quantitative indicators or large-scale surveys, providing limited insight into how organizations actually experience and implement digital transformation in practice. As a result, the processual and context-specific nature of digital adoption remains underexplored.

This study addresses these limitations by providing a qualitative, empirically grounded analysis of digital technology and AI adoption in Slovak companies. By integrating technological, organizational, and human perspectives, the study contributes to a more comprehensive understanding of digital transformation as a multi-dimensional and context-dependent process.

Table I presents selected indicators from the Digital Economy and Society Index (DESI), highlighting differences between Slovakia and the European Union average in key areas of digital skills. While Slovakia performs comparably in basic digital skills and exceeds the EU average in digital content creation, it lags behind in more advanced indicators, such as the share of individuals with above-basic digital skills, the proportion of ICT specialists, and the level of enterprise engagement in employing ICT professionals.

TABLE I. COMPARISON OF DIGITAL SKILLS IN SLOVAKIA AND THE EUROPEAN UNION IN 2022

Digital skills	DESI	
	DESI Slovakia	DESI EU
At least basic digital skills (% of population)	55	54
Above basic digital skills (% of population)	21	26
At least basic skills in digital content creation (% of population)	72	66
ICT specialists (% of employed persons aged 15–74)	4.30	4.50
Female ICT specialists (% of ICT specialists)	15	19
Enterprises employing ICT specialists (% of enterprises)	16	20
ICT graduates (% of graduates)	4.40	3.90

Source: [9]

These disparities point to structural limitations in the country's digital readiness, particularly in terms of advanced competencies and workforce specialization. As suggested by the literature, such gaps may constrain the ability of firms to adopt and effectively utilize more sophisticated digital technologies, including artificial intelligence, which require not only technical infrastructure but also highly skilled human capital [23], [30].

Beyond national-level indicators, digital transformation is increasingly understood as a multidimensional process that extends beyond technological capability. Prior research emphasizes the importance of organizational culture, leadership support, and strategic alignment in shaping digital transformation outcomes [17], [32]. In this context, employee reactions represent a critical factor influencing implementation success. Resistance to technological change is a well-documented phenomenon, particularly among employees with lower levels of digital skills or longer-established work routines [31], [24].

However, the literature also suggests that such resistance can be mitigated through effective change management practices, including communication, employee involvement, and training initiatives [19]. At the same time, emerging research highlights the role of continuous learning and skill development as key mechanisms enabling employees to adapt to rapidly evolving technological environments.

Digital transformation also has broader implications for labor markets and employment structures. While automation and AI may substitute routine tasks, studies increasingly indicate that digitalization simultaneously generates new roles requiring higher-order cognitive and technical skills [1], [11], [14]. This dual effect reinforces the importance of ongoing upskilling and organizational investment in human capital as essential components of sustainable digital transformation. At the same time, recent research points to the existence of a "productivity paradox" in AI adoption, whereby the implementation of AI technologies does not immediately translate into measurable productivity gains at the aggregate level [22]. Instead, performance improvements tend to emerge over time, as firms

invest in complementary organizational changes, workforce skills, and intangible assets. This suggests that the benefits of AI adoption are inherently long-term and dependent on broader organizational transformation processes.

Taken together, the literature suggests that digital transformation is best understood as an interaction between three interrelated dimensions: technological innovation, strategic organizational change, and human-capital development. This integrated perspective provides a relevant analytical foundation for examining how Slovak companies perceive and implement digital technologies within their specific institutional and economic context.

III. Methodology

This study employs a qualitative research design to explore how Slovak companies perceive and implement digitalization and digital transformation. Because the research questions are exploratory in nature and aim to capture subjective experiences, meanings and organizational contexts, a qualitative approach based on semi-structured interviews was selected. Such approaches are particularly suitable for examining complex and dynamic processes that are socially and contextually embedded [7], and are widely used in exploratory research where in-depth understanding of organizational phenomena is required.

A. Research Design

This study adopts a qualitative research design to explore how Slovak companies perceive and implement digitalization and artificial intelligence (AI). Given the exploratory nature of the research questions and the focus on understanding organizational processes, meanings, and experiences, a qualitative approach based on semi-structured interviews was considered most appropriate [6], [15]. This design enables an in-depth examination of complex and context-dependent phenomena that cannot be adequately captured through standardized quantitative methods.

B. Research Questions

The study is guided by the following research questions:

- RQ1: How do Slovak companies perceive digitalization and digital transformation?
- RQ2: How do companies adopt and use artificial intelligence and other digital tools?
- RQ3: How do employees react to digital changes and how do they adapt?
- RQ4: What organizational effects does digitalization produce in Slovak companies?

C. Sampling and Participants

A purposive sampling strategy was employed to select information-rich cases across different organizational contexts [26]. The sample consists of 20 Slovak companies representing a range of sectors, including information technology, finance, manufacturing, education, and services, as well as different firm sizes (micro, small, medium, and large enterprises).

The selection criteria included:

- Active use of digital technologies in organizational processes,
- Willingness to participate in the study, and
- Representation of sectoral diversity.

Although the sample size is relatively modest, it is consistent with qualitative research standards, where adequacy is determined by thematic saturation rather than statistical representativeness. In this study, thematic saturation was achieved after approximately 15 interviews, with subsequent interviews confirming previously identified patterns. This approach is consistent with qualitative research standards, where analytical depth and contextual richness are prioritized over statistical generalization, and where smaller samples are considered appropriate for capturing complex organizational processes [6], [25].

TABLE II. Structure of the Research Sample by Sector and Firm Size

Sector/Company size	Micro (1-9)	Small (10-49)	Medium (50-249)	Large (more than 250)	Total
IT	3	0	2	1	6
Finance	1	1	0	2	4
Education	2	0	0	0	2
Business services	0	1	0	0	1
Culture	0	1	0	0	1
Gastronomy	0	0	1	0	1
Manufacturing	0	0	1	1	2
Retail/Wholesale	0	0	0	1	1
Law and Legislation	0	0	0	1	1
Marketing	0	0	0	1	1
Total	6	3	4	7	20

Source: Authors' own processing

Table II summarizes the composition of the research sample, capturing the distribution of participating companies across sectors and firm sizes. The sample reflects a deliberate effort to ensure heterogeneity, including micro, small, medium, and large enterprises operating in diverse industries such as information technology, finance, manufacturing, education, and services.

This variation was intentionally incorporated to capture a wide range of organizational experiences with digitalization and AI adoption, as prior research indicates that sectoral context and firm size significantly influence the scope and depth of technological implementation. In particular, the inclusion of both technologically intensive sectors (e.g., IT, manufacturing) and less digitally advanced sectors (e.g., culture, education) enables a comparative perspective on digital maturity and adoption patterns.

Although the sample size (N = 20) is relatively modest, it is consistent with qualitative research standards, where the objective is not statistical representativeness but analytical depth and contextual understanding. In line with established methodological approaches [6], [25], sampling adequacy was assessed based on thematic saturation. In this study, saturation was reached after approximately fifteen interviews, with subsequent interviews providing confirmatory rather than novel insights.

This sampling strategy strengthens the study's ability to identify recurring patterns while preserving sensitivity to sector-specific differences, thereby enhancing the analytical robustness and transferability of the findings.

D. Data Collection

Data were collected through semi-structured interviews conducted between March and July 2025. Two interview guides were developed to capture both managerial and employee perspectives:

A manager-focused guide addressing digital strategy, AI adoption, organizational change, and digital competencies.

An employee-focused guide examining individual experiences, perceived challenges, and adaptation processes.

Interviews lasted between 30 and 60 minutes and were conducted either in person or online. All interviews were audio-recorded with participants' consent and subsequently transcribed verbatim for analysis.

E. Data Analysis

The data were analyzed using thematic analysis following the six-step framework proposed by [3]:

- Familiarization with the data.
- Generation of initial codes.
- Search for themes.
- Review of themes.
- Definition and naming of themes.
- Production of the report.

Thematic analysis is particularly suitable for identifying patterns across qualitative data and for developing theoretically informed interpretations grounded in empirical evidence, making it an appropriate method for the exploratory objectives of this study.

A hybrid inductive–deductive approach was applied. Deductive codes were derived from the research questions (e.g., "AI adoption," "digital skills," "organizational change"), while inductive codes emerged from participants' responses (e.g., "initial resistance," "informal learning," "automation benefits").

To enhance analytical transparency, a codebook was developed specifying code definitions and representative examples. Coding was conducted by the primary researcher and partially reviewed by a second researcher through peer debriefing, which contributed to the consistency and credibility of the analysis.

F. Analytic Framework (Conceptual Formalization)

Based on the thematic analysis, the study develops a conceptual framework of digital technology and AI adoption. The framework can be formally represented as a three-level structure of organizational digital maturity:

Level 1 (L_1): Basic digitization – use of digital tools for administrative support and data management.

Level 2 (L_2): Process automation – integration of digital systems for workflow optimization and process coordination.

Level 3 (L_3): AI-supported adoption – implementation of AI-based tools for decision support and advanced analytics.

These levels are influenced by three cross-cutting dimensions:

- Employee adaptation (E),
- Organizational restructuring (O),
- Sectoral context (S).

The relationship between these components can be conceptually expressed as:

$$D = f(L, E, O, S) \quad (1)$$

where D represents the level of digital transformation achieved by the organization, L denotes the stage of technological adoption, and E, O, and S represent moderating organizational and contextual factors. This formalization serves as a conceptual abstraction of the relationships identified through qualitative analysis, providing a structured representation of how technological, organizational, and human factors jointly shape digital transformation outcomes.

This formalization does not imply a quantitative model but serves as an analytical representation of relationships identified through qualitative data.

G. Trustworthiness and Validity

To ensure the credibility and trustworthiness of the findings, several strategies were applied:

Triangulation through the inclusion of both managerial and employee perspectives.

Thick description using illustrative quotations to provide contextual depth.

Audit trail documenting coding decisions and analytical steps.

Peer debriefing to validate selected coding procedures. These procedures enhance the credibility, consistency, and analytical rigor of the findings, in line with established qualitative research standards.

H. Ethical Considerations

The study adhered to standard ethical principles, including informed consent, voluntary participation, anonymity, and confidentiality. All participants were informed about the purpose of the study, and identifying information was removed during data processing.

I. Limitations

As a qualitative pilot study, the findings are not statistically generalizable. The sample is based on voluntary participation, which may introduce self-selection bias. However, the study provides analytically transferable insights that can inform future research using larger or mixed-methods designs.

IV. Results

The findings of the study are presented according to the four research questions, with themes identified through thematic analysis illustrating common patterns and divergent experiences across organizations. Overall, the analysis revealed three dominant patterns across organizations: a predominantly process-oriented understanding of digitalization, significant sectoral differences in the adoption of advanced digital technologies, and gradual employee adaptation following initial resistance to technological change. The thematic analysis revealed four dominant themes across organizations (see Table III).

TABLE III. Overview of Themes Identified Through Thematic Analysis

Theme	Description	Illustrative evidence
Process-oriented understanding of digitalization	Companies primarily perceive digitalization as a tool for improving efficiency and automating administrative processes.	Managers frequently describe digitalization as automation of administrative tasks or workflow optimization.
Sectoral differences in technology adoption	The level of technological adoption varies across sectors, with IT and manufacturing firms demonstrating higher digital maturity.	Companies in IT and manufacturing reported using cloud systems, data analytics and automation tools.
Employee resistance and gradual adaptation	Employees initially express resistance to technological change but adapt over time through practice-based learning.	One respondent noted that changes in work systems initially caused discomfort but became easier once benefits were recognized.
Organizational consequences of digitalization	Digitalization leads to job redesign, task reallocation and efficiency improvements, while workforce reductions are relatively rare.	Several managers reported reassigning employees to new roles rather than eliminating positions.

Source: Authors' own processing

A. Process-Oriented Understanding of Digitalization

Addressing RQ1, companies generally perceived digitalization as a process-oriented initiative focused on improving workflow efficiency and reducing manual labor. Managers frequently described digitalization in technical terms, such as automating administrative procedures or optimizing communication systems, while a smaller number articulated a broader understanding of digital transformation as organizational change. This variation suggests inconsistent conceptual clarity across firms. As one retail manager noted, the introduction of digital tools "streamlined several administrative processes, reducing the amount of manual work, particularly in logistics." Although efficiency was widely emphasized, few firms linked digital transformation to employee learning or strategic skill development, indicating a primarily technology-driven approach.

B. Sectoral Differences in Digital Technology Adoption

Regarding RQ2, the adoption of advanced digital technologies, including artificial intelligence, varied across sectors. IT and manufacturing companies reported the greatest technological integration, such as cloud services, data analytics tools, and automation systems. AI use remained limited and exploratory, typically focused on chatbots or automated support functions. Companies in education, culture, law, and business services relied mainly on basic digital tools without substantial transformation initiatives. Organizational resources and size influenced the depth of adoption, with larger firms investing in complex systems such as SCADA or HMI panels, whereas micro and small enterprises pursued incremental and cost-sensitive solutions.

C. Employee Resistance and Gradual Adaptation

In relation to RQ3, employees' reactions to digital changes followed a common pattern: initial resistance, followed by gradual adaptation. Respondents described hesitation, discomfort, and uncertainty during the early stages of technological implementation, a reaction especially pronounced among older workers. An IT employee (aged 56) explained, "I felt resistance. Every change in the work system causes discomfort. It is important to anticipate this and design an implementation strategy in advance." Despite this resistance, employees generally adapted over time as they recognized the benefits of reduced manual tasks and clearer processes. Adaptation occurred through practice rather than through structured training, indicating an experiential learning process.

In several cases, respondents emphasized that the adaptation process was facilitated by informal peer support and gradual exposure to digital tools in everyday work activities. Rather than relying on formal training programs, many organizations introduced new technologies incrementally, allowing employees to learn through repeated use and interaction with colleagues. Managers reported that this approach reduced anxiety associated with technological change and helped employees build confidence when working with digital systems. In smaller firms in particular, learning often occurred through collaborative problem solving, where employees shared knowledge and practical tips regarding the use of new software or digital platforms. These findings suggest that workplace-based learning mechanisms play an important role in supporting employees' adaptation to digitalization, especially in organizations with limited resources for structured training initiatives.

D. Organizational Consequences of Digitalization

Addressing RQ4, digitalization also produced structural effects within organizations. While some firms experienced changes in job roles and responsibilities, workforce reductions were uncommon. Managers commonly reported reassigning employees to new positions rather than eliminating roles. A finance manager stated. "Digitalization has not affected the number of employees. No positions were eliminated, but the people in those positions have changed." In certain sectors, however, digitalization prompted outsourcing or automation

that replaced specific operational tasks. For example, a manufacturing manager described outsourcing IT and analytics functions to partners abroad and implementing SCADA and HMI technologies that "eliminated several operational positions two years ago." Across organizations, digitalization consistently contributed to improved efficiency, workflow clarity, and time savings, with employees often redirected to higher-value tasks.

E. A Framework of Digital Technology and AI Adoption

Building on the thematic analysis, the findings of this study allow for the development of a conceptual framework that captures how companies adopt digital technologies and artificial intelligence across different stages of organizational development.

The analysis identifies three distinct levels of digital adoption. The first level, basic digitization, involves the use of digital tools primarily for administrative support and data management. At this stage, digitalization is understood mainly as a means of improving efficiency and reducing manual workload, which corresponds with the process-oriented understanding identified across most organizations.

The second level, process automation, is characterized by the integration of more advanced digital systems, including cloud-based platforms, workflow automation tools, and data-processing technologies. Companies at this stage move beyond basic digitization and begin to restructure internal processes to enhance operational efficiency and coordination.

The third level, AI-supported adoption, represents a more advanced stage in which organizations experiment with or implement artificial intelligence technologies, such as chatbots, automated analytics, or decision-support tools. However, the findings indicate that this level remains limited and unevenly distributed across sectors.

Across all levels, the adoption process is shaped by several cross-cutting dimensions. First, employee adaptation plays a crucial role, as employees typically exhibit initial resistance to technological change followed by gradual adaptation through practice-based learning. Second, digitalization contributes to organizational restructuring, particularly through task reallocation and job redesign rather than large-scale workforce reductions. Third, sectoral differences significantly influence the depth of adoption, with technologically intensive industries demonstrating higher levels of digital maturity.

Taken together, the framework suggests that digital transformation is not a linear technological upgrade but a multi-dimensional process combining technological implementation with human and organizational adaptation.

V. Discussion and Implications

The findings of this study provide important insights into how Slovak companies approach digitalization and artificial intelligence (AI) adoption, while also contributing to broader debates on digital transformation in organizational contexts. Importantly, the results highlight that digital transformation should be understood not merely as a technological upgrade, but as a multi-dimensional process shaped by the interaction of technological, organizational, and human factors. Taken together, the findings suggest that digital transformation in Slovak firms follows a gradual and non-linear trajectory shaped by both internal organizational capabilities and external contextual constraints.

A. Reinterpreting Digitalization from Operational Tool to Strategic Process

The results indicate that Slovak companies predominantly conceptualize digitalization as a process-oriented tool aimed at improving efficiency and automating administrative tasks. This finding aligns with prior research suggesting that organizations often adopt digital technologies in a reactive and operational manner before recognizing their broader strategic implications [4], [28]. This finding contributes to the ongoing debate by suggesting that the transition from operational digitalization to strategic digital transformation is not automatic, but depends on the alignment of technological investments with organizational capabilities and human capital development.

However, the limited distinction observed between digitalization and digital transformation suggests a lack of strategic integration, particularly in smaller firms. While the literature emphasizes that digital transformation requires alignment between technology, strategy, and organizational capabilities [27], the findings indicate that many firms remain at an early stage of this transition. This reinforces the argument that technological adoption alone does not constitute transformation but must be embedded within a broader organizational change process.

B. Sectoral Inequalities and Uneven Technological Maturity

Consistent with existing research, the study reveals significant sectoral differences in digital adoption. Firms operating in IT and manufacturing sectors demonstrate higher levels of technological maturity, including the use of automation, cloud systems, and early-stage AI applications. This pattern reflects findings by [11], [5], who highlight the role of resource availability, technological capabilities, and innovation intensity in shaping digital transformation trajectories.

In contrast, firms in education, culture, and service-oriented sectors exhibit slower and more incremental adoption. These differences can be interpreted through the lens of resource constraints and perceived necessity, which have been identified as key barriers to digital transformation in SMEs. The findings therefore support the view that digital transformation is uneven and context-dependent, rather than a uniform or linear process across industries.

C. Employee Adaptation from Resistance to Experiential Learning

A central contribution of this study lies in its identification of employee adaptation as a dynamic and processual phenomenon. Consistent with established models of technology acceptance [31], [24], employees initially exhibit resistance to technological change, particularly when faced with unfamiliar systems or disrupted routines.

However, the findings extend this literature by demonstrating that resistance is typically temporary and followed by gradual adaptation through practice-based learning. Rather than relying on formal training programs, many organizations facilitate adaptation through informal

mechanisms, including peer support, collaborative problem-solving, and repeated interaction with digital tools. This aligns with the concept of workplace-embedded learning [7], highlighting the importance of experiential learning in digital transformation processes.

At the same time, the reliance on informal learning may limit the long-term development of digital competencies, suggesting that organizations could benefit from more structured approaches to skill development. This observation contributes to the ongoing debate on the role of human capital in digital transformation, reinforcing the need to integrate technological and educational strategies. These findings extend existing models of technology acceptance by emphasizing the role of informal and practice-based learning as a key mechanism of adaptation, particularly in resource-constrained organizational environments.

D. Organizational Consequences: Transformation Without Displacement

The findings suggest that digitalization primarily leads to job redesign, task reallocation, and efficiency gains rather than widespread workforce reductions [1], [2], [21]. This is consistent with prior research indicating that digital transformation often reshapes rather than eliminates work [1], [2].

Nevertheless, evidence of outsourcing and automation in certain sectors, particularly manufacturing, indicates that digitalization can also contribute to structural shifts in employment. This dual effect reflects broader labor market dynamics, where digital technologies simultaneously create new opportunities while displacing specific routine tasks. The findings therefore underscore the importance of adaptive organizational strategies that support workforce transition and skill development. This suggests that the impact of digitalization on employment should be understood as a process of transformation rather than substitution, where the nature of work changes more significantly than the number of jobs.

E. Contextualizing Digital Transformation in Slovakia

While many of the observed patterns align with broader European trends, the findings also highlight several context-specific characteristics of the Slovak environment. Digital transformation appears to be shaped by a combination of structural constraints, including limited resources in SMEs, gaps in advanced digital skills, and reliance on external expertise through outsourcing.

These factors correspond with macro-level indicators such as DESI, which point to moderate digital readiness but persistent gaps in workforce capabilities and enterprise-level adoption. Importantly, the study demonstrates how these structural conditions are reflected at the firm level, influencing both the pace and depth of digital transformation. This contextual perspective contributes to a more nuanced understanding of digitalization in CEE economies, where institutional and economic conditions differ from more technologically advanced regions.

F. Contributions, Limitations, and Implications

This study contributes to the literature by offering an empirically grounded and conceptually integrated perspective on digital transformation in a CEE context. Specifically, it provides three main contributions. First, it provides empirically grounded insights into digital and AI adoption in a CEE context, addressing a gap in existing literature. Second, it conceptualizes digital transformation as a multi-level process, distinguishing between basic digitization, process automation, and AI-supported adoption. Third, it highlights the critical role of employee adaptation and organizational learning in shaping transformation outcomes.

At the same time, several limitations should be acknowledged. The study is based on a relatively small and non-random sample, which limits the generalizability of the findings. In addition, the reliance on qualitative interviews may introduce subjective bias in participants' responses. Future research could address these limitations by employing mixed methods designs, larger samples, and longitudinal approaches to capture the evolution of digital transformation over time.

From a practical perspective, the findings suggest that organizations should approach digital transformation as an integrated process combining technological investment with human-capital development and organizational change. Policymakers may also play a role by supporting SME digitalization, strengthening digital skills development, and fostering collaboration between industry and educational institutions.

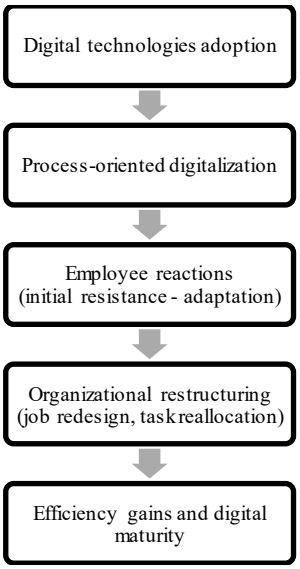


Fig. 1. Conceptual framework of digital technology and AI adoption, Source: Authors' own elaboration based on qualitative data.

Fig. 1 presents the conceptual framework of digital technology and AI adoption derived from the thematic analysis. The framework illustrates three progressive levels of digital adoption—basic digitization, process automation, and AI-supported adoption—reflecting increasing levels of technological maturity within organizations.

In addition to these stages, the model highlights three cross-cutting dimensions that shape the adoption process: employee adaptation, organizational restructuring, and sectoral context. These dimensions influence both the pace and the depth of digital transformation, emphasizing that technological implementation is closely intertwined with human and organizational factors.

The framework conceptualizes digital transformation as a dynamic and multi-dimensional process rather than a linear progression, where organizations may move between stages depending on internal capabilities and external conditions. This perspective extends existing models of digital transformation by explicitly integrating employee-level adaptation and sector-specific variability into a unified analytical structure.

VI. Conclusion and Practical Implications

This study examined how Slovak companies perceive, adopt, and respond to digitalization and artificial intelligence (AI), drawing on qualitative evidence from 20 organizations across multiple sectors. The findings demonstrate that digitalization remains predominantly process-oriented, with firms focusing on workflow optimization and incremental technological improvements rather than comprehensive strategic transformation. While some organizations recognize the broader implications of digital transformation, the integration of human capital development into digital strategies remains limited. These findings reinforce the view that digital transformation in SMEs is not a purely technological process, but a socio-technical transition requiring alignment between technology, organization, and human capabilities.

The analysis further reveals that digital adoption is uneven across sectors, with IT and manufacturing firms exhibiting higher levels of technological maturity, while service-oriented sectors adopt digital tools more gradually and primarily for operational purposes. AI adoption remains exploratory and limited in scope, reflecting both resource constraints and uncertainty regarding its practical implementation. At the organizational level, digitalization leads primarily to task reallocation, job redesign, and efficiency gains, rather than widespread workforce reductions.

A key contribution of the study is the development of a conceptual framework that conceptualizes digital transformation as a multi-level process comprising basic digitization, process automation, and AI-supported adoption. Importantly, the findings highlight that this process is not purely technological but is shaped by employee adaptation, organizational restructuring, and sectoral context. This integrated perspective extends existing literature by explicitly linking technological adoption with human and organizational dimensions.

From a practical perspective, the findings suggest that organizations should approach digital transformation as a structured and phased process. In the initial stage, firms should focus on implementing basic digital tools to improve data management and reduce manual workload. In the second stage, emphasis should shift toward process automation and the integration of digital systems to enhance efficiency and coordination. Finally, in the advanced stage, organizations may begin to experiment with AI-based solutions, provided that sufficient technological and human capabilities are in place.

At the same time, the study underscores the importance of supporting employee adaptation through continuous learning, knowledge sharing, and effective change management. Without adequate investment in human capital, the potential benefits of digital technologies may remain underutilized. These findings also have implications for policymakers, particularly in the context of Central and Eastern Europe, where targeted support for SME digitalization, digital skills development, and collaboration between academia and industry may accelerate digital transformation.

The study is subject to several limitations. The relatively small and non-random sample limits the generalizability of the findings, and the qualitative design may be influenced by subjective interpretations. Future research could address these limitations by employing mixed-methods approaches, larger datasets, and longitudinal designs to capture the evolution of digital transformation processes over time. Comparative studies across countries or sectors would further enhance understanding of contextual differences in digital adoption.

Acknowledgment

This paper is an output of the research project funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project of the Smart Transformational and Innovation Consortium Slovakia (STICS), No. 09I0203V0100011, by VEGA project no. 1/0551/24: Basic Research into New Paradigms of Social Marketing in the Context of Post-Pandemic Restructuring of Societal Values, by Slovak Research and Development Agency (SRDA), grant no. APVV-24-0607: Development of Multi-dimensional Models for Assessing Corporate AI Readiness and Adoption Related to the Digital Transformation Strategy of Slovakia 2030.

References

[1] D. Autor, D. Mindell, and E. Reynolds, The Work of the Future: Building Better Jobs in an Age of Intelligent Machines. MIT Press, 2022.

[2] M. Arntz, T. Gregory, and U. Zierahn, "The risk of automation for jobs in OECD countries: A comparative analysis," OECD Social, Employment and Migration Working Papers, no. 189, OECD Publishing, 2016.

[3] V. Braun and V. Clarke, "Using thematic analysis in psychology," Qualitative Research in Psychology, vol. 3, no. 2, pp. 77–101, 2006.

[4] S. Brennen and D. Kreiss, "Digitalization," in The International Encyclopedia of Communication Theory and Philosophy, Wiley, 2016.

[5] M. G. Colombo, E. Piva, and C. Rossi-Lamastra, "Digitalization and innovation in SMEs: A review of the literature," Small Business Economics, vol. 58, no. 2, pp. 669–700, 2022.

[6] J. W. Creswell and C. N. Poth, Qualitative Inquiry and Research Design: Choosing Among Five Approaches, 4th ed. SAGE, 2018.

[7] M. Eraut, "Informal learning in the workplace," Studies in Continuing Education, vol. 26, no. 2, pp. 247–273, 2004.

[8] European Commission, "Digital Economy and Society Index (DESI) 2022: Thematic chapters," 2022.

[9] Eurostat, "Digitalization in Europe – 2023 edition," 2023.

[10] J. Fähndrich, "A literature review on the impact of digitalization on management control," Journal of Management Control, vol. 34, no. 1, pp. 9–65, 2023.

[11] A. G. Frank, L. S. Dalenogare, and N. F. Ayala, "Industry 4.0 technologies: Implementation patterns in manufacturing companies," International Journal of Production Economics, vol. 210, pp. 15–26, 2019.

[12] J. Huňady, "Drivers of AI adoption in SMEs: Missing skills and other firm-level factors," *Journal of Decision Analytics and Intelligent Computing*, vol. 5, no. 1, pp. 99–110, 2025. doi:10.31181/jdaic10023062025h.

[13] J. Huňady, Ž. Lacová, and A. Vallušová, "AI adoption by European companies: Drivers and barriers in the EU," in G. Pellat, Ž. Lacová, and M. Šuplata, Eds., *Data-Centric Business and Applications: How a Business and Territorial Approach Can Challenge the Crises in Europe?*, vol. 258. Springer, 2025, pp. 339–348. doi:10.1007/978-3-031-95068-1_17.

[14] E. Ivanová, E. Koišová, and J. Masárová, "Zmeny v štruktúre trhu práce v kontexte rozvoja digitálnej ekonomiky (Changes in the structure of the labour market in the context of digital economy development)," in *Vplyv Industry 4.0 na tvorbu pracovných miest*, Košice: TUKE, 2021, pp. 189–198.

[15] H. Kallio, A. M. Pietilä, M. Johnson, and M. Kangasniemi, "Developing a framework for a qualitative semi-structured interview guide," Journal of Advanced Nursing, vol. 72, no. 12, pp. 2954–2965, 2016.

[16] G. C. Kane, D. Palmer, A. N. Phillips, D. Kiron, and N. Buckley, "Strategy, not technology, drives digital transformation," MIT Sloan Management Review, 2015.

[17] R. Kohli and N. P. Melville, "Digital innovation: A review and synthesis," Information Systems Journal, vol. 29, no. 1, pp. 200–228, 2019.

[18] J. Kádárová, M. Fil'o, D. Sukopová, and M. Dúlová, "Artificial intelligence adoption and labour productivity in Slovakia and the EU27: Implications for sustainable economic growth," *Sustainability*, vol. 18, 2026.

[19] J. P. Kotter, Leading Change. Harvard Business Review Press, 2012.

[20] J. Kunecová, J. Prester, J. Šebo, and I. Palčič, "Why do manufacturing firms struggle with artificial intelligence?" *Journal of Manufacturing Technology Management*, 2025. doi:10.1108/JMTM-05-2025-0432.

[21] T. Masárová, "Faktory digitalizácie ekonomiky a ich vplyv na zamestnania," (*Factors of digitalization of the economy and their impact on employment*). Revue sociálno-ekonomického rozvoja, vol. 4, no. 1, pp. 87–93, 2020.

[22] MIRRI SR, "Index digitálnej ekonomiky a spoločnosti (DESI)," 2024.

[23] OECD, OECD Skills Outlook 2019: Thriving in a Digital World. OECD Publishing, 2019.

[24] S. Ransbotham, D. Kiron, P. Gerbert, and M. Reeves, "Reshaping business with artificial intelligence," *MIT Sloan Management Review*, vol. 59, no. 1, pp. 1–17, 2017.

[25] S. Oreg, J. M. Bartunek, G. Lee, and B. Do, "An affective model of organizational change," The Journal of Applied Behavioral Science, vol. 47, no. 1, pp. 33–62, 2011.

[26] M. Q. Patton, Qualitative Research & Evaluation Methods, 4th ed. SAGE, 2015.

[27] J. Reis, M. Amorim, N. Melão, and P. Matos, "Digital transformation: A literature review and guidelines for future research," Springer, 2018.

[28] T. Tóth and M. Hallová, "Artificial intelligence in enterprises: A comparative analysis of Slovakia and EU member states," in *International Scientific Days 2024*, pp. 429–437, 2024. doi:10.15414/2024.9788055228167.429-437.

[29] D. R. A. Schallmo and C. A. Williams, "History of digital transformation," in Digital Transformation Now!, Springer, 2018.

[30] E. Van Laar, A. J. A. M. van Deursen, J. A. G. M. van Dijk, and J. de Haan, "Determinants of 21st-century skills and digital skills for workers," Computers in Human Behavior, vol. 112, 2020.

[31] J. Veber and J. Srpová, Digitalizace ekonomiky a společnosti (Digitalization of the economy and society), Praha: Management Press, 2018.

[32] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," MIS Quarterly, vol. 27, no. 3, pp. 425–478, 2003.

[33] P. C. Verhoef et al., "Digital transformation: A multidisciplinary reflection and research agenda," Journal of Business Research, vol. 122, pp. 889–901, 2021.

[34] G. Vial, "Understanding digital transformation: A review and a research agenda," MIS Quarterly, vol. 43, no. 1, pp. 121–165, 2019.

Feature-Level Analysis and Robust Baselines for EEG-Based Imagined Speech Recognition on the ASU Dataset

Hatem T M Duhair[1], Masrullizam Mat Ibrahim[2]*,
Jamil Abedalrahim Jamil Alsayaydeh[3]*, Mazen Farid[4]*, Safarudin Gazali Herawan[5]
Department of Engineering Technology-Fakulti Teknologi Dan Kejuruteraan Elektronik Dan Komputer (FTKEK), Universiti Teknikal Malaysia Melaka, Melaka, Malaysia[1, 2, 3]
Faculty of Information Science and Technology (FIST), Multimedia University, Melaka 75450, Malaysia[4]
Centre for Intelligent Cloud Computing, COE for Advanced Cloud, Multimedia University, Melaka 75450, Malaysia[4]
Faculty of Engineering-Industrial Engineering Department, Bina Nusantara University, Jakarta, Indonesia[5]

***Abstract*—Imagined speech decoding from non-invasive electroencephalography remains a challenging problem, especially when moving beyond small vocabularies and optimistic evaluation protocols. This work revisits the Arizona State University (ASU) imagined speech dataset and treats it as a rigorous ten-class benchmark, with a focus on offline, corpus-level analysis rather than real-time deployment. After unifying all recordings into 5 s epochs at 256 Hz, 6,520 trials with 60 EEG channels were preprocessed using bandpass filtering, baseline correction, z-score normalization, and trial-wise ICA for artifact attenuation. On top of this pipeline, a comprehensive feature representation was constructed that combines common spatial patterns, discrete wavelet statistics, time-domain moments, autocorrelation coefficients, power spectral density band powers, and Hjorth parameters into a single 5,120-dimensional vector. A block-wise ablation indicates that autocorrelation, CSP, PSD, and Hjorth features carry most of the discriminative information in this setting, while wavelet and simple statistical descriptors contribute little and can be removed without harming performance. Using only the informative blocks (3,440 features), a multinomial logistic regression classifier reaches about 0.41 accuracy and 0.42 macro F1 on the ten-class task, roughly four times chance level. A multi-layer perceptron and a CNN–LSTM model, trained under the same splits and with class weighting, do not outperform this linear baseline and exhibit stronger overfitting. Within the evaluated protocol, these findings suggest that carefully engineered features capture most of the discriminative structure accessible on this corpus, and that deeper models add complexity without clear benefit. The study provides a transparent baseline and a feature-level analysis that can serve as a reference point for future work on imagined speech recognition and transfer learning across EEG corpora.**

Keywords*—EEG-based imagined speech; ASU imagined speech dataset; brain–computer interface; feature extraction; logistic regression

I. Introduction

Electroencephalography (EEG) based imagined speech recognition has emerged as a promising route toward silent speech interfaces for users who cannot rely on the vocal tract, for example people with advanced neuromuscular disease or severe dysarthria. By decoding internal speech representations directly from cortical activity, such systems aim to restore basic communication with minimal motor effort and without external articulation. Recent surveys on silent speech interfaces and EEG-based brain computer interfaces consistently highlight imagined speech as one of the most challenging yet societally important application domains [1], [2].

Despite rapid advances in deep learning, the performance of noninvasive imagined speech decoders remains modest when evaluated in realistic conditions. Most published studies operate on highly constrained tasks, such as binary vowel classification or discrimination between a few short words, and often use subject-specific models with carefully tuned pipelines. Reported accuracies in these settings typically lie around 70 to 80 percent, but they tend to fall toward 20 to 50 percent once the vocabulary is expanded, the class structure becomes more heterogeneous, or evaluation protocols adopt stricter cross-validation schemes [3], [4]. This gap between optimistic benchmark numbers and robust multiclass performance has motivated calls for more systematic baselines and ablation studies that clarify what can realistically be achieved with current sensors, preprocessing strategies, and feature representations.

The Arizona State University (ASU) imagined speech corpus is one of the few publicly available datasets that support such analysis [5]. It contains multi-session EEG recordings of short words, longer phrases and isolated vowels, acquired at 256 Hz with a high-density montage and segmented around the final auditory cue [6]. Compared with smaller imagined speech datasets, ASU introduces several sources of difficulty that are frequently encountered in real use: multiple lexical categories, variable trial lengths, and imbalanced class distributions that reflect differences in the number of repetitions per condition. Previous work on ASU has mainly focused on end-to-end architectures that combine time frequency representations with convolutional or recurrent neural networks, usually reporting single model accuracies without detailed inspection of the underlying feature space [7].

In parallel, the broader EEG community has developed a rich toolbox of handcrafted feature extractors, such as common spatial patterns, wavelet coefficients, spectral band powers and Hjorth parameters, that have proved effective in motor imagery

*Corresponding authors

and other BCI paradigms [2]. However, their behaviour in the context of imagined speech is less well documented. Many studies adopt a fixed subset of features without quantifying how each block contributes to classification, whether different descriptors are redundant or complementary, or how simple linear classifiers compare with more elaborate deep architectures when both are applied to the same feature space. This lack of systematic feature-level analysis makes it difficult to interpret negative results, to compare methods across papers, or to design compact models that might eventually run in real time on portable hardware.

The present work addresses these gaps through a deliberately transparent study of EEG-based imagined speech recognition on the ASU dataset. A reproducible preprocessing pipeline is first constructed that performs channel selection, bandpass filtering, baseline correction, trial length normalization and artifact removal using independent component analysis. The cleaned signals are then transformed into a large unified feature set... On top of this representation, several strong but interpretable classifiers are evaluated, including multinomial logistic regression, linear support vector machines, multilayer perceptrons and a lightweight CNN LSTM network, using consistent train validation test splits and macro-averaged metrics to account for label imbalance. The intended deployment scenario is explicitly offline: the goal is to characterise what handcrafted EEG features can achieve on the ASU corpus under a fixed evaluation protocol, rather than to prototype a real-time BCI. The resulting baselines are framed as a diagnostic reference that any subsequent real-time system, with its tighter latency and subject-adaptation constraints, would have to match or exceed.

Beyond reporting headline accuracies, particular emphasis is placed on understanding which feature families actually support discrimination in this ten-class imagined speech task. A leave-one-block-out analysis and reduced feature experiments quantify the marginal contribution of each feature group and show, perhaps counterintuitively, that a smaller subset that combines spatial filters, autocorrelation, band power, and Hjorth parameters can match or slightly improve the performance of the full feature set. These results provide a nuanced picture: overall macro F1 scores remain in the range of 0.40 to 0.44, which confirms the intrinsic difficulty of the problem on this dataset, yet the systematic ablations reveal meaningful structure that can inform subsequent model design.

In summary, this study makes three main contributions. First, it offers a carefully documented preprocessing and feature extraction pipeline for the ASU imagined speech corpus that other groups can reuse as a common baseline. Second, it provides a comparative evaluation of classical and deep classifiers on a unified feature space, highlighting where additional model complexity does and does not translate into improved generalization. Third, it delivers a detailed feature-level analysis that clarifies which descriptors carry the most useful information for ten-class imagined speech recognition. Taken together, these contributions position transparent baselines as an essential complement to more ambitious end-to-end architectures, helping the field move toward more realistic expectations and more robust future systems.

The remainder of this study is organised as follows: Section II reviews related work on EEG-based imagined speech recognition, with particular attention to the ASU corpus and to feature engineering practices in non-invasive brain–computer interfaces. Section III describes the proposed methodology, including the preprocessing pipeline, the multi block feature extraction scheme, and the classification models and evaluation protocols. Section IV presents the experimental results on the ten-class ASU task, reports the feature ablation and model comparison findings, and discusses their implications for future imagined speech systems. Finally, Section V summarises the main conclusions, highlights the limitations of the current study, and outlines several directions for extending the work through cross-dataset transfer learning and more compact real-time implementations.

II. Related Works

Research on EEG-based imagined speech recognition has expanded notably in the last five years, but performance remains far below that of motor imagery or overt speech decoding. Recent systematic reviews show that most EEG speech imagery systems are evaluated on small, highly constrained vocabularies and often under subject-dependent protocols, where models are trained and tested on the same individual. Under such conditions, reported accuracies can exceed 80 %, yet they typically fall to the 40–60 % range when the vocabulary grows, when multi-class settings are used, or when cross-subject generalisation is considered [8]. This gap between optimistic, laboratory-scale results and more realistic, multi-class performance is central to the present work.

A first line of recent studies still relies heavily on hand-crafted features combined with relatively shallow classifiers. Before surveying individual contributions, it is useful to fix the three axes along which they differ most and that determine whether a reported number is directly comparable to another: 1) the scale and paradigm of the dataset, including the number of participants, the number of imagined classes, and whether the corpus uses short vowels, isolated words, phrases or continuous material; 2) the evaluation protocol, spanning subject-dependent versus cross-subject splits, stratified trial-level versus session-level partitioning, and whether macro-averaged metrics are used in the presence of class imbalance; and 3) the model class, ranging from linear projections on engineered features to convolutional, recurrent or hybrid deep architectures operating on raw or minimally processed EEG. Headline accuracies that look superficially similar can correspond to very different tasks along these axes, and most of the apparent disagreement in the literature dissolves once studies are stacked against this grid. The studies surveyed below are ordered accordingly, moving from small-vocabulary subject-dependent pipelines with shallow classifiers toward multi-class, cross-subject deep models, and are referenced back to this frame whenever they are introduced. These works exploit time-frequency decompositions, power spectral descriptors, and spatial projections such as common spatial patterns in order to enhance signal-to-noise ratio before classification. For example, deep metric learning has been used to map EEG trials for five vowels and six words into a latent space where distances reflect phonetic

similarity, achieving around 45 % accuracy in a six-word imagined speech task despite strong inter-subject variability [9]. Other studies report comparable performance when combining wavelet-based features with traditional machine learning models, particularly when the number of commands is limited, and the recording protocol is tightly controlled [6]. These results suggest that carefully engineered features can extract some discriminative structure from EEG speech imagery, but they do not fully resolve the problem of class overlap in larger label spaces.

More recent work has shifted towards deep learning architectures that operate directly on raw or minimally processed EEG. Abdulghani et al. used wavelet scattering representations of 8-channel EEG and an LSTM network to classify four imagined directional commands, reporting an accuracy of about 92.5 % under a subject-specific setting, which illustrates how recurrent models can capture temporal dynamics when the task is well constrained [10]. In contrast, when the vocabulary increases and multi-class decoding is attempted, accuracy tends to decline. Alharbi et al. proposed hybrid 3D CNN-StackLSTM and 3D CNN-BiLSTM models that operate on sequences of EEG topographic maps for five imagined words in the BCI2020 dataset. Their best model reached an average accuracy of roughly 45 %, with substantial variability across subjects, even though the architecture explicitly encodes both spatial and temporal information [11]. Similar hybrid CNN-BiLSTM systems, often combined with wavelet-based front-ends or metaheuristic feature selection, have shown strong single-subject performance but less convincing cross-subject generalisation [12].

Another strand of work explores transfer learning and representation learning to mitigate the scarcity of labelled imagined speech data. Lee et al. applied deep metric learning to jointly optimise a feature space for vowels and words, enabling reuse of representations across tasks and datasets and achieving mid-40 % accuracy for six-class problems [9]. Other authors have investigated transfer learning from overt speech or from related EEG paradigms. One recent study proposed a transfer learning framework that jointly extracts temporal and spectral features, then fine-tunes a classifier on limited imagined speech trials, showing consistent but modest gains over training from scratch [13], [14]. These approaches underline that cross-task and cross-corpus transfer is feasible, yet the improvements are incremental and far from closing the gap to practical communication rates.

The limitations of current performance are closely tied to data availability. For many years, researchers relied on small institutional datasets with a handful of commands or vowels, which restricted the diversity of linguistic material and impeded robust evaluation. The recent release of the Chinese Imagined Speech Corpus (Chisco), which contains more than 20,000 sentences of high-density EEG per subject, is a major step forward and offers a much richer testbed for neural language decoding from EEG [15]. However, most published work with Chisco and similar corpora is still at an early stage and often focuses on feasibility demonstrations rather than systematic baselines across feature families and model classes [8].

Against this backdrop, the present study takes a deliberately conservative and diagnostic perspective. Instead of aiming for a complex, highly tuned end-to-end model, the analysis is performed on a medium-scale imagined speech dataset (ASU) that contains ten classes and 6,520 trials, using a rich but interpretable feature space that combines common spatial patterns, discrete wavelet features, statistical descriptors, autocorrelation, band-power estimates, and Hjorth parameters. The discriminative contribution of each feature family is then quantified by analysing block-wise performance with linear models and by comparing these baselines to compact neural architectures such as multilayer perceptrons and CNN-LSTM hybrids. The results, which cluster around 40–44 % macro F1 for ten-class classification, are consistent with the more demanding multi-class findings in the recent literature and provide a transparent reference point for future work on transfer learning and model compression in EEG-based imagined speech recognition.

III. Materials and Methods

This section describes the complete processing pipeline used in this study, from raw electroencephalographic (EEG) recordings to the classification of imagined speech commands. The workflow is organised into three tiers. Tier 1 covers data acquisition from the ASU imagined speech corpus and the preprocessing applied to the 60-channel EEG signals, including filtering, baseline correction, artefact rejection, and quality control. Tier 2 details the feature engineering stage, in which six complementary feature blocks are extracted and combined into a reduced 3,440-dimensional representation. Tier 3 introduces the discriminative models and evaluation procedures that operate on this compact feature space. An overview of these stages and their dependencies is summarised in Fig. 1.

A. Dataset and Labelling

The ASU imagined speech dataset was collected by the Human Oriented Robotics and Controls (HORC) laboratory at Arizona State University using a 64-channel BrainProducts ActiCHamp amplifier. Recordings were acquired from 15 healthy volunteers (11 males and 4 females), who were instructed to silently rehearse visually and auditorily cued prompts without overt articulation or muscle movements. The vocabulary comprised four groups of stimuli: short commands (“in”, “out”, “up”), long words (“cooperate”, “independent”), short–long combinations (“in”, “cooperate”), and sustained vowels (/a/, /i/, /u/). Each trial followed a structured cueing protocol that aligned the onset of imagined speech relative to a sequence of auditory beeps, a design that is now standard in imagined speech paradigms [5], [16].

This study focused on the "last-beep" interval, that is, the segment immediately preceding the final cue, which has been reported to concentrate the most stable imagined speech activity while limiting contamination from anticipatory and post-movement processes [16]. After aggregating all usable files and retaining only last-beep epochs, a total of $N = 6520$ trials distributed across ten imagined classes. These classes correspond to three short commands (“in”, “up”, “out”), two long words (“cooperate”, “independent”), two short–long phrase conditions (“cooperate”, “in”), and three vowels (“a”, “i”, “u”). For supervised learning, each semantic class was mapped to an

integer label $y \in \{0, \ldots, 9\}$, following a consistent mapping used throughout the subsequent analysis and model training.

The original montage contained dedicated electrooculography (EOG) channels that are useful for artefact detection but not for decoding imagined speech. For each trial, the designated EOG indices (0, 9, 32, and 63 in the original layout) were removed, and the first $C = 60$ EEG channels as the working montage. All signals were sampled at $f_s = 256$ Hz, and trial lengths varied slightly across recordings because of minor differences in cue timing and file boundaries. Before feature extraction and modelling, these variable-length epochs were transformed into a fixed-size tensor through the preprocessing pipeline described in Section III.B. The chosen epoch length of approximately 5 s corresponds to the full last-beep interval defined by the ASU acquisition protocol and used in previous ASU-based analyses; it encompasses the imagined articulation itself together with a short post-cue window during which sustained speech-motor activity has been reported. Shorter windows would truncate this sustained activity and eliminate the lag range exploited by the autocorrelation and Hjorth descriptors introduced later, whereas substantially longer windows would reintroduce pre-cue anticipation and post-trial drift that the last-beep alignment was designed to avoid.

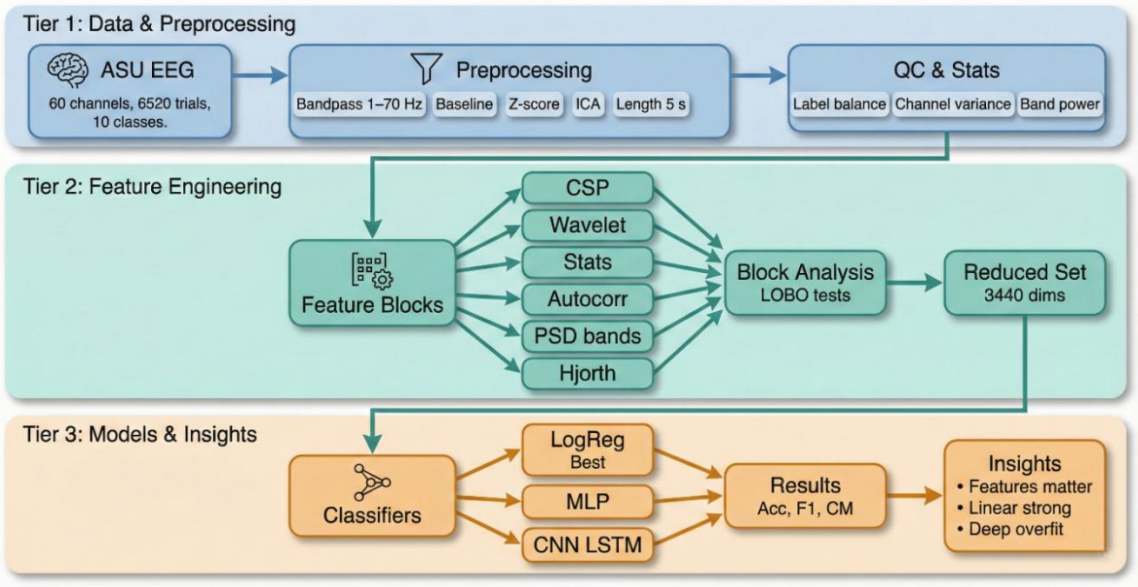


Fig. 1. Overview of the proposed EEG imagined speech recognition pipeline on the ASU dataset

B. Preprocessing Pipeline

All preprocessing was implemented in Python using MNE-Python together with standard scientific libraries (NumPy, SciPy, scikit-learn). The design goal was to obtain artefact-reduced and temporally aligned trials while preserving the spectral content that is typically informative for imagined speech and BCI tasks [17]. The same sequence of operations was applied to every trial: temporal alignment by zero-padding or truncation, band-pass filtering, baseline correction, trial-wise z-score normalization, and independent component analysis (ICA) for artefact attenuation. Fig. 2 displays the raw versus the preprocessed data for the dataset using our pipeline.

Let $x_c(t)$ denote the raw signal from channel $c \in \{1, \ldots, C\}$ at discrete time index t, with trial-specific length T. First, a common temporal length $T_{\max}$, defined as the maximum number of samples across all trials in the dataset. For trials shorter than $T_{\max}$ we applied zero-padding at the end, and for trials longer than $T_{\max}$ we truncated the surplus samples:

$$\bar{x}_c(t) = \begin{cases} x_c(t), & 0 \le t < T, \\ 0, & T \le t < T_{max}, \end{cases} \quad (1)$$

This operation produces a temporally aligned representation without interpolation, which helps preserve the phase relationships of oscillatory activity that are often exploited in EEG decoding [18]. At this stage, each trial has shape $C \times T_{\max}$.

To suppress slow drifts and high-frequency sensor noise while retaining the classical delta to low-gamma range relevant for cognitive and speech-related processes, a zero-phase fifth-order Butterworth band-pass filter with passband $[f_{\text{low}}, f_{\text{high}}] = [1,70]$ Hz. The filter was implemented using forward-backward IIR filtering (filtfilt) to avoid phase distortion [19]. For each channel, the band-passed signal is:

$$y_c(t) = \mathcal{F}_{\text{BP}}\left(\bar{x}_c(t); f_{\lambda o \omega} = 1 \text{ Hz}, f_{\eta \iota \gamma \eta} = 70 \text{ Hz}\right), \quad (2)$$

where, F_{BP} denotes the zero-phase band-pass operator applied along the time axis.

Slow baseline shifts were then removed at the trial level. For the ASU dataset, a dedicated baseline interval of 200 ms was available at the beginning of each epoch. Let B be the set of baseline time indices, corresponding to [0,0.2]s at 256 Hz. The baseline mean for the channel c is:

$$\mu_c^{\text{base}} = \frac{1}{|B|} \sum_{t \in B} y_c(t). \quad (3)$$

The baseline-corrected signal is then:

$$z_c(t) = y_c(t) - \mu_c^{\beta \alpha \sigma \varepsilon}. \quad (4)$$

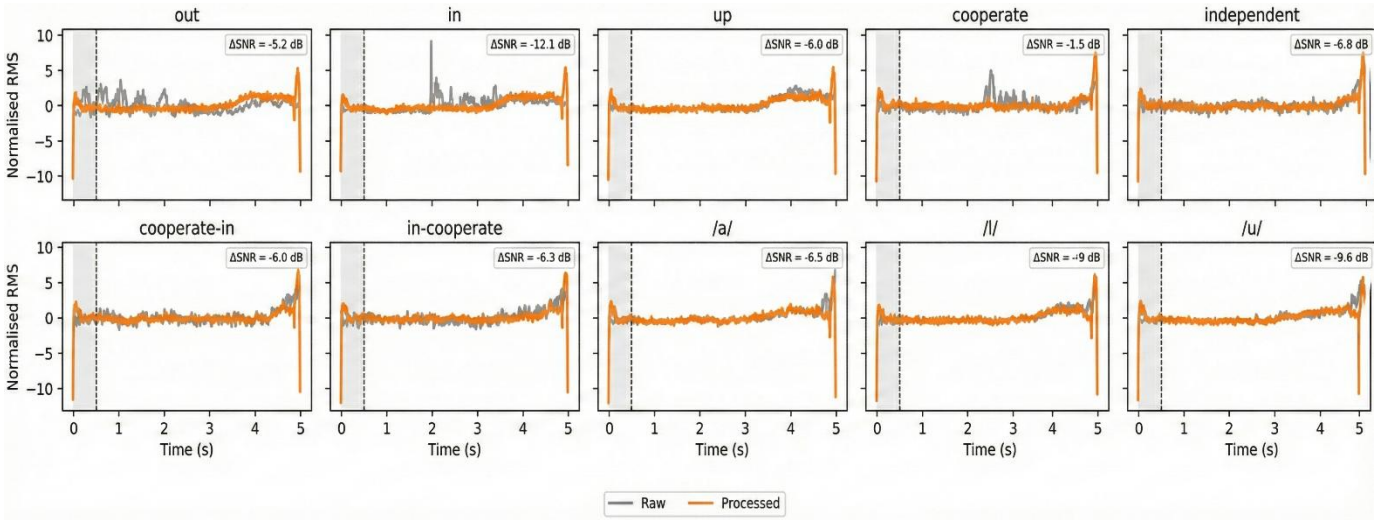


Fig. 2. Raw versus processed classes of the ASU dataset.

When a dedicated baseline window is not explicitly defined, the first 5% of the trial can be used as an approximate baseline, which is a common practical choice in continuous EEG paradigms [20].

To stabilise subsequent covariance-based methods and ensure that channels contribute on comparable scales, trial-wise z-score normalization was performed per channel. For each channel c, the mean and standard deviation of the baseline-corrected signal were computed over the entire trial,

$$\mu_c = \frac{1}{T_{max}} \sum_{t=0}^{T_{max}-1} z_c(t), \tag{5}$$

$$\sigma_c = \sqrt{\frac{1}{T_{max}} \sum_{t=0}^{T_{max}-1} (z_c(t) - \mu_c)^2}, \tag{6}$$

and obtained the normalized time series:

$$\hat{x}_c(t) = \frac{z_c(t) - \mu_c}{\sigma_c + \varepsilon} \tag{7}$$

where, ε is a small constant used for numerical stability. This per-trial normalization mimics the standardisation strategies widely adopted in deep EEG models and large-scale BCI datasets [21], [22].

Residual eye movements and muscle artefacts were attenuated using ICA. For each trial, an MNE RawArray was constructed with $C = 60$ channels and sampling rate $f_s = 256\ Hz$, and estimated an ICA decomposition with as many components as channels. Let $X \in \mathbb{R}^{C \times T_{\max}}$ denote the normalized multichannel signal. ICA models X as a linear mixture of statistically independent sources:

$$X = AS \tag{8}$$

where, $A \in \mathbb{R}^{C \times C}$ is the mixing matrix and $S \in \mathbb{R}^{C \times T_{max}}$ contains the independent components. Components whose time courses and topographies resembled canonical ocular or myogenic artefacts were excluded heuristically, following established guidelines for EEG artefact handling [23]. The cleaned signal is then reconstructed as:

$$\hat{X} = A\hat{S} \tag{9}$$

where, $\hat{S}$ equals S with the identified artefact components set to zero. ICA-based correction remains one of the most effective approaches for denoising EEG in both clinical and BCI contexts [24].

After preprocessing, each trial is represented by a cleaned tensor $\hat{X} \in \mathbb{R}^{C \times T_{\max}}$. Stacking all $N = 6520$ trials yields the final preprocessed dataset:

$$\hat{\mathcal{X}} \in \mathbb{R}^{N \times C \times T_{max}}, \tag{11}$$

which serves as the input to the subsequent feature extraction and classification stages described in the following subsections.

C. Feature Extraction

We designed a unified feature extraction block that compresses each preprocessed trial into a 5120-dimensional vector. The block combines six complementary descriptors: spatial filters from Common Spatial Patterns (CSP), time frequency statistics from a discrete wavelet transform, low-order time domain moments, lagged autocorrelation, band power from power spectral density (PSD), and Hjorth mobility and complexity. Similar hybrid designs have been shown to outperform single-family features in motor imagery and speech imagery BCIs, since they capture both transient and sustained dynamics across multiple scales. The resulting representation is deliberately broad rather than minimal: the intent is to expose redundancy between feature families so that the block-wise ablation in Section IV-B can quantify which descriptors genuinely contribute and which can be removed. All features were computed per trial and concatenated, then normalised once at the end using a single StandardScaler fitted on the training partition, avoiding information leakage into the test set. The trade-off between feature breadth and redundancy is examined directly in the ablation results, and the reduced 3,440-dimensional subset is presented there as the operational representation used by the classifiers [25].

1) Common spatial patterns: To capture spatially discriminative activity across electrodes, CSP was applied in a one versus rest setting for each imagined class. Let denote a single trial with channels and time samples, and let and be the

class averaged covariance matrices for the target class and the pooled background, respectively. For each binary problem CSP solves the generalised eigenvalue problem $X \in \mathbb{R}^{C\times T}$ $C T \Sigma_1 \Sigma_2$:

$$\Sigma_1 \mathbf{w}_k = \lambda_k \Sigma_2 \mathbf{w}_k, \quad (12)$$

where, $\mathbf{w}_k$ is the k-th spatial filter and λ_k the associated eigenvalue. Projecting the trial as $\mathbf{z}_k = \mathbf{w}_k^T X$ and computing the log-normalised variance:

$$f_k^{\mathrm{CSP}} = \log \frac{\mathrm{var}(\mathbf{z}_k)}{\sum_j \mathrm{var}(\mathbf{z}_j)} \quad (13)$$

yields CSP features that emphasise spatial patterns with maximal class-specific variance [26]. The implementation allocates n_{comp} filters per class, chosen adaptively from the number of channels and classes, which in the ASU configuration results in 60 CSP features per trial.

2) Wavelet-based time frequency statistics: Fine-grained time frequency structure was represented using a Level 3 Discrete Wavelet Transform (DWT) with a Daubechies 4 Mother Wavelet. Db4 is frequently adopted in EEG work because it provides compact support and smooth oscillatory atoms that align well with dominant EEG rhythms in the alpha and beta ranges [6]. For each channel, the preprocessed time series is decomposed into one approximation and three detail coefficient vectors. For each coefficient vector of length six, statistics are computed: $c, x_c(t)\{\mathbf{c}_\ell\}_{\ell=0}^{3} n$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} c_i, \quad (14)$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n} (c_i - \mu)^2 \quad (15)$$

$$\mathrm{skew} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{c_i-\mu}{\sigma}\right)^3, \quad (16)$$

$$\mathrm{kurt} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{c_i-\mu}{\sigma}\right)^4, \quad (17)$$

along with the minimum and maximum of c. Concatenating these statistics across all levels and 60 channels yields a 1440-dimensional DWT feature block that captures transient, band-limited changes associated with imagined speech cues [27].

3) Time domain statistics: To summarise coarse distributional properties in the time domain, four low-order moments are computed directly from the baseline corrected and normalised signal. For each channel, we estimate the mean, standard deviation, skewness, and kurtosis over time and concatenate them across channels. The resulting 240-dimensional block provides a compact characterisation of overall amplitude, variability, and waveform asymmetry. Although simple, such statistics have been reported to add complementary information when combined with more structured features in EEG classification tasks [28] $cx_c(t)$.

4) Autocorrelation features: Imagined speech tends to recruit sustained, quasi-rhythmic activity that extends over several hundred milliseconds, and articulatory-motor planning signals have been reported to unfold over comparable timescales in overt and covert speech paradigms. To capture these temporal dependencies, the one-sided autocorrelation is computed for each channel up to a maximum lag samples, which corresponds to approximately 0.2 s at 256 Hz. This horizon is short enough to avoid washing out the moment-to-moment dynamics of a 5 s trial, yet long enough to span the dominant alpha/beta cycles (roughly 50–400 Ms) that are implicated in speech-motor and sensorimotor integration, and it is comparable to the autocorrelation ranges adopted in earlier imagined-speech and motor-imagery feature studies. For channel, $\tau_{\max} = 50c$:

$$r_c(\tau) = \sum_{t=0}^{T_{\max}-1-\tau} x_c(t)x_c(t+\tau), \tau = 0, \ldots, \tau_{\max} \quad (18)$$

The first 50 autocorrelation coefficients per channel are concatenated to form a 3000-dimensional block across all 60 channels. These features encode the decay and periodicity of channel-wise dynamics without imposing stationarity assumptions and have been shown to benefit EEG-based decoding of cognitive and motor states [29].

5) Power spectral density and band power: Frequency-specific information is extracted using Welch's Method for PSD Estimation. Each channel is segmented into overlapping windows, tapered with a Hamming window, and the periodograms are averaged to obtain. The PSD is then integrated within five canonical bands $PSD_c(f)$:

$$P_c^{\theta} = \sum_{f\in[4,8)} PSD_c(f), \quad (19)$$

$$P_c^{\beta} = \sum_{f\in[13,30)} PSD_c(f), \quad (20)$$

$$P_c^{\delta} = \sum_{f\in[0.5,4)} PSD_c(f), \quad (21)$$

$$P_c^{\alpha} = \sum_{f\in[8,13)} PSD_c(f), \quad (22)$$

$$P_c^{\gamma} = \sum_{f\geq 30} PSD_c(f), \quad (23)$$

yielding five band powers per channel and 300 PSD-based features per trial. Band power in these ranges remains a workhorse descriptor in motor imagery and speech imagery BCIs and offers a robust, interpretable link to underlying oscillatory processes [30].

6) Hjorth parameters: Finally, Hjorth mobility and complexity parameters are computed, which summarise the spectral content of a signal using simple variance ratios in the time domain. For channel, let denote the time series, its first discrete derivative, and its second derivative. Mobility and complexity are defined as $cx_c(t)\dot{x}_c(t)\ddot{x}_c(t)$:

$$\mathrm{mobility}_c = \sqrt{\frac{\mathrm{var}(\dot{x}_c(t))}{\mathrm{var}(x_c(t))}}, \quad (24)$$

$$\mathrm{complexity}_c = \sqrt{\frac{\mathrm{var}(\ddot{x}_c(t))}{\mathrm{var}(\dot{x}_c(t))}} / \mathrm{mobility}_c. \quad (25)$$

These quantities reflect the dominant frequency and the change in frequency over time and have been repeatedly validated as efficient EEG features for BCI and clinical applications [31]. Concatenating mobility and complexity

across the 60 channels produces an 80-dimensional Hjorth feature block.

7) Feature normalisation and reduced subset: All six blocks are concatenated to form a feature matrix, where trials. A single StandardScaler is then fitted on to obtain normalised features with zero mean and unit variance per dimension, which are used by both linear and deep models. To reduce redundancy and computational cost, we also define a reduced subset by retaining only the CSP, Autocorrelation, PSD and Hjorth Blocks, resulting in a 3440-dimensional representation. As discussed later in the results, this subset preserves most of the discriminative structure in a simple Multinomial Logistic Regression Analysis, while substantially lowering the input dimensionality for the neural classifiers. $F \in \mathbb{R}^{N\times 5120} N = 6520 F\tilde{F}$.

D. Model Architectures and Training Procedure

All classification experiments on the ASU imagined speech dataset were carried out on the reduced feature representation described in Section III-C. Each trial is represented by a 3440-dimensional vector $\mathbf{x} \in \mathbb{R}^{3440}$, obtained by concatenating spatial (CSP), temporal (autocorrelation), spectral (PSD), and dynamical (Hjorth) descriptors from the preprocessed EEG. The ASU corpus contributes 6520 labelled trials with ten classes ($y \in \{0, \ldots, 9\}$), corresponding to short, long, short-long, and vowel commands. This setting allows us to decouple model behaviour from low-level signal processing and to focus on how different architectures exploit the same feature space, in line with recent work on deep learning for EEG-based BCIs.

1) Baseline and deep architecture: Three discriminative models were considered: a Multinomial Logistic Regression Baseline, a Multilayer Perceptron (MLP), and a Deeper CNN-BiLSTM-FCNN Hybrid. The motivation for this choice was comparative rather than architectural optimisation. The linear model establishes an interpretable reference in the same feature space. The MLP tests whether a modest amount of non-linear recombination of the engineered features improves discrimination. The CNN-BiLSTM-FCNN is included as a deliberate stress test: although the 3,440-dimensional feature vector is a heterogeneous concatenation of spatial, spectral and temporal descriptors and therefore lacks the strict local ordering that convolutions typically exploit in raw EEG decoding, this architecture has been reported to improve performance on pre-engineered EEG features in several imagined-speech studies. Including it on the same representation allows an empirical test of whether any residual sequential structure along the feature axis can be recovered by a deeper sequence model, and if so, at what computational cost. Alternative architectures that are better suited to tabular inputs – for instance, gradient-boosted ensembles or attention-based tabular encoders – are considered promising directions for follow-up work but are not evaluated here, so as to preserve a controlled three-way comparison on a common feature space. All three implement the same probabilistic mapping via a softmax output layer, but differ in how they transform the input features. The detailed architectures of the three classifiers are illustrated in Fig. 3: $p_\theta(y \mid \mathbf{x})$.

For the logistic regression baseline, the conditional class probabilities are given by:

$$p_\theta(y = k \mid \mathbf{x}) = \frac{\exp\left(\mathrm{w}_k^\top \mathrm{x} + b_k\right)}{\sum_{j=0}^{C-1} \exp\left(\mathrm{w}_j^\top \mathrm{x} + b_j\right)}, k = 0, \ldots, C-1, \quad (26)$$

with $C = 10$ and $\theta = \{\mathrm{w}_k, b_k\}_{k=0}^{C-1}$. This model provides a well-understood linear baseline to contextualise the gains from non-linear feature transformations.

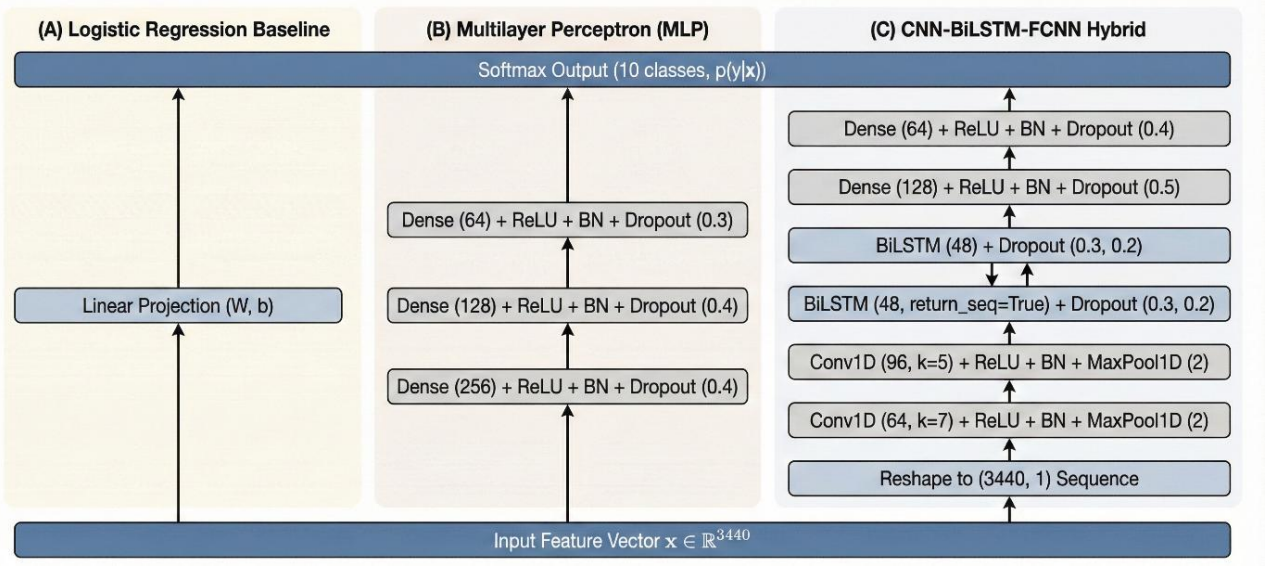


Fig. 3. Model architectures used in the ASU experiments: (A) Logistic regression baseline, (B) Multilayer perceptron with three hidden layers, and (C) CNN–BiLSTM–FCNN hybrid operating on the 3 440-dimensional feature sequence.

The improved MLP operates directly on x and stacks three fully connected hidden layers with batch normalisation and dropout. If $\mathrm{h}^{(0)} = \mathrm{x}$, each hidden layer l computes

$$\mathrm{h}^{(l)} = \phi\left(\mathrm{W}^{(l)}\mathrm{h}^{(l-1)} + \mathrm{b}^{(l)}\right), \tag{27}$$

Where, $\phi(\cdot)$ is the rectified linear unit (ReLU). The final softmax layer produces $\hat{\mathbf{y}} = \text{softmax}\left(\mathrm{W}^{(L)}\mathrm{h}^{(L-1)} + \mathrm{b}^{(L)}\right)$. All dense layers are regularised with an ℓ_2 penalty on the weights to reduce overfitting in this relatively high-dimensional space. Such compact MLPs have been found competitive on engineered EEG features when the sample size is modest.

The CNN-BiLSTM-FCNN model treats each feature vector as a 1D sequence of length 3440 with a single channel and therefore seeks local patterns along the feature axis. Two 1D convolutional blocks first extract local motifs:

$$\mathbf{z}_t^{(c)} = \phi\left(\sum_j \sum_\tau w_{j,\tau}^{(c)} x_{j,t+\tau} + b^{(c)}\right), \tag{28}$$

followed by batch normalisation and max pooling to build increasingly abstract representations. The resulting sequence of feature maps is fed to stacked bidirectional LSTM layers that capture longer range dependencies by processing the sequence in both forward and backward directions. Bidirectional recurrent architectures are known to improve sequence modelling in speech and EEG applications by exploiting past and future context. The final temporal embedding is passed through a small fully connected head with ReLU, batch normalisation, and dropout before the softmax classifier.

Table I summarises the main architectural choices. All hyperparameters are fixed across experiments on the ASU dataset in order to attribute performance differences to the model families rather than to extensive tuning.

TABLE I. TRAINING AND OPTIMIZATION SETTINGS FOR THE THREE CLASSIFIERS ON THE REDUCED ASU FEATURE SET

Model	Input shape	Optimizer	Initial LR	Max epochs	Batch size	Early stopping	LR schedule	Regularisation/class weighting
Logistic regression	(3440)	L2-regularised multinomial solver	–	–	–	–	–	ℓ_2 penalty on **W** class_weight = balanced
MLP (deep baseline)	(3440)	Adam	3×10−4	150	64	val_loss, 25	factor 0.5, patience 8, min_lr 10^{-5}	ℓ_2 on all dense layers, batch norm, dropout 0.4–0.3
CNN-BiLSTM-FCNN hybrid	(3440,1)	Adam	3×10−4	120	64	val_loss, 20	factor 0.5, patience 6, min_lr 10^{-5}	ℓ_2 on FC layers, batch norm, dropout 0.5–0.4, class_weight = balanced

2) Loss function, class weighting, and optimization: All models are trained with a weighted categorical cross-entropy loss. Given a training set, the objective is $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^N w_{y_i} \log\, p_\theta(y_i \mid \mathbf{x}_i), \tag{29}$$

where, w_{y_i} is the class weight associated with label y_i. Class weights are computed inversely proportional to class frequencies in the training partition, which mitigates the moderate label imbalance observed in the ASU corpus and avoids the need for synthetic oversampling. Similar class reweighting strategies are standard practice in EEG classification when some commands appear less frequently than others.

As illustrated in Table I, all neural models use the Adam optimiser with an initial learning rate of 3×10^{-4}. Training proceeds for a maximum of 150 epochs for the MLP and 120 epochs for the CNN-BiLSTM-FCNN, using mini-batches of size 64. To reduce manual tuning, a Reduce-on-Plateau scheduler halves the learning rate whenever the validation loss fails to improve for several epochs (8 epochs for the MLP, 6 for the CNN-BiLSTM-FCNN), with a floor at 10^{-5}. Early stopping with patience of 25 epochs (MLP) and 20 epochs (CNN-BiLSTM-FCNN) restores the weights that achieve the lowest validation loss, which helps to prevent overfitting and stabilises the reported results.

3) Data partitioning and evaluation metrics: For each model, the 6520 ASU Trials are split with stratification into training, validation, and held-out test data. Stratification preserves the empirical label distribution across splits and ensures that rare classes are represented in both validation and test sets. No subject-specific information is used in this stage, since the goal of this experiment is to characterise how well different architectures exploit the engineered feature space rather than to probe cross-subject generalisation $60\%, 20\%, 20\%$.

The data were split at the trial level into train, validation, and test partitions of 60 %, 20 %, and 20 %, respectively, preserving the mild class imbalance described in Section III. To report performance in a way that is robust to this imbalance, overall accuracy and macro-averaged F1 are the primary metrics of interest. All models were trained with a single fixed random seed to keep the three-way comparison on strictly identical splits; the resulting point estimates should therefore be read as single-run performance on the held-out test set rather than as means of a sampling distribution. A formal repeated-runs analysis with confidence intervals over multiple seeds and resampled stratified splits is a natural next step and is discussed among the study's limitations. Accuracy is defined as:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{\hat{y}_i = y_i\} \tag{30}$$

while macro-F1 averages per-class F1 scores,

$$F_1^{\text{macro}} = \frac{1}{K}\sum_{k=1}^K \frac{2\Pi\rho\varepsilon\chi\iota\sigma\iota\sigma\nu_k \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}, \tag{31}$$

with $K = 10$ classes. Macro-F1 is particularly informative for imagined speech where some commands are intrinsically harder to discriminate than others. This metric treats all classes equally, which is crucial in imagined speech tasks where misclassifying a rarely used command can be more problematic

than misclassifying a frequent one. Confusion matrices and full per class precision, recall, and F1 scores are additionally reported to reveal systematic confusions between short, long, and vowel commands and to support later qualitative analysis of error patterns.

Taken together, this design yields a controlled comparison between a linear baseline, a compact deep MLP, and a more expressive CNN–BiLSTM–FCNN architecture on exactly the same ASU feature representation.

IV. Results and Discussion

A. Overall Classification Performance

After feature engineering and block-wise reduction, the ASU imagined speech dataset yielded a compact representation of 6,520 trials in a 3,440-dimensional feature space with ten classes (three vowels, three phrase variants, and four directional commands). Table II summarises the test performance of the three classifiers evaluated on the reduced feature set.

The numbers are consistent with the bar plot in Fig. 4, which juxtaposes accuracy and macro-F1 for each model.

The first observation is that the linear logistic baseline already reaches a macro-F1 of about 0.42, essentially matching the deeper MLP. This confirms that once the EEG is projected into a carefully designed feature space, simple discriminative models can remain highly competitive, a pattern also reported for CSP-based motor imagery and imagined speech pipelines [32], [33]. The MLP does not significantly improve aggregate metrics, although it offers slightly more stable learning dynamics and provides a natural stepping stone toward later transfer learning stages in our broader programme.

The CNN–BiLSTM–FCNN hybrid, in contrast, underperforms slightly despite its far higher computational cost. For the MLP, both training and validation loss decrease rapidly during the first 30 epochs, after which validation loss plateaus while training loss continues to fall. Validation accuracy stabilises around 0.42, whereas training accuracy climbs above 0.60, indicating mild overfitting but no catastrophic divergence. A similar pattern is present, but more pronounced, for the CNN–BiLSTM model, whose validation loss flattens early while the network continues to specialise on the training set. These dynamics suggest that the reduced feature set does not provide enough independent evidence for the deeper temporal model to exploit, a finding in line with recent reports that end-to-end deep networks for non-invasive imagined speech often saturate near 30–40 % accuracy on vocabularies of 6–11 words [34], [35].

The confusion matrix in Fig. 5 further clarifies where errors arise. Vowel classes /a/, /i/, and /u/ show relatively high diagonal terms, with many trials correctly localised, while the directional commands “out”, “in”, and “up” are frequently confused with one another. The social-interaction phrases “cooperate”, “independent”, “cooperate-in”, and “in-cooperate” occupy an intermediate regime, with good separation between “cooperate-in” and the rest, but noticeable cross-talk between “cooperate” and “independent”. This structure mirrors previous findings that imagined vowels tend to elicit more stereotyped sensorimotor and auditory patterns than semantically richer phrases.

TABLE II. Test Accuracy, Macro-F1, and Qualitative Computational Cost for Three Classifiers Trained on the Reduced ASU Feature Set

Model	Test Accuracy (%)	Test macro-F1 (%)	Qualitative cost profile*
Logistic regression	41	42	Very low, single linear layer.
MLP (three-layer network)	41	42	Moderate, ≈ 60 epochs, fast per epoch.
CNN–BiLSTM–FCNN hybrid	40	38	High, long epochs and large parameter count.
*Cost profile is based on empirical training times and model size, not on formal complexity analysis.			

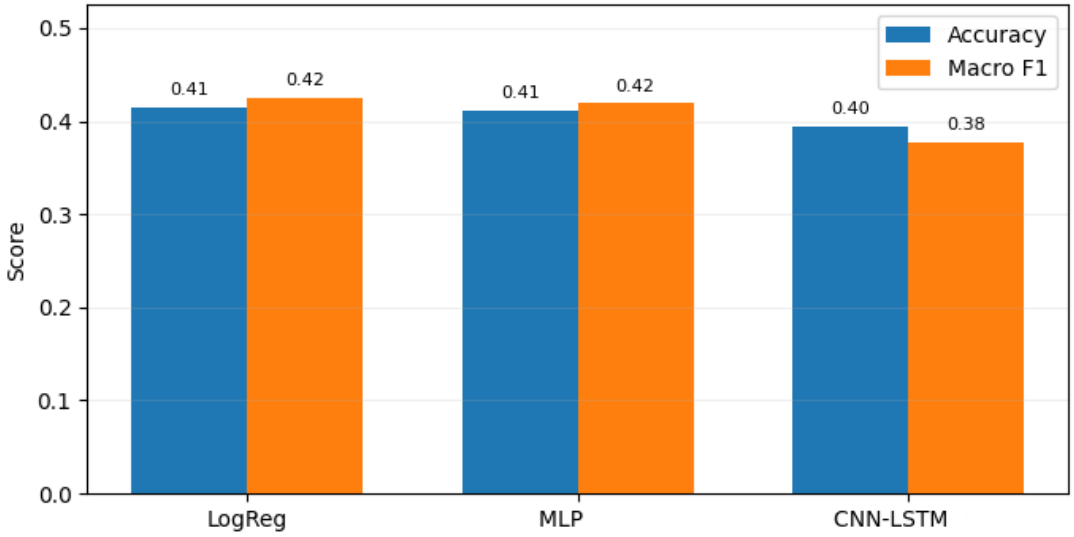


Fig. 4. Comparison of test accuracy and macro-F1 for logistic regression, MLP, and CNN–BiLSTM–FCNN on the reduced ASU features.

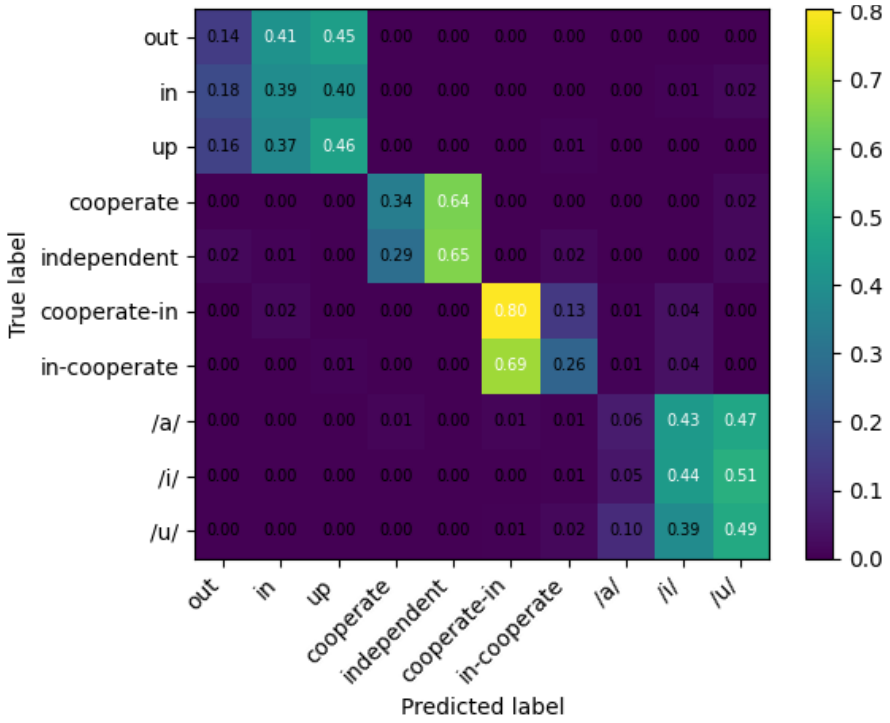


Fig. 5. Normalised confusion matrix for the CNN–BiLSTM–FCNN classifier on the ASU test set, showing systematic confusions among directional commands and vowels.

B. *Block-wise Contribution of Engineered Feature Families*

To understand how much each feature family contributes to the final performance, we conducted two complementary ablation studies on the logistic regression classifier: 1) single-block runs where the model receives only one family of features, and 2) leave-one-block-out (LOBO) runs where a given block is removed from the full feature set. The four blocks considered here are CSP spatial filters, autocorrelation statistics, band-power features from the power spectral density (PSD), and Hjorth parameters capturing activity, mobility, and complexity. These blocks have a long history in EEG decoding and are known to capture partially complementary aspects of oscillatory dynamics and temporal structure.

Table III summarises the macro-F1 obtained in the single-block and LOBO conditions, which correspond to the bar plots in Fig. 6.

TABLE III. BLOCK-WISE MACRO-F1 SCORES WITH ONE FEATURE FAMILY AND CHANGE IN MACRO-F1 WHEN EACH BLOCK IS REMOVED (LEAVE-ONE-BLOCK-OUT ANALYSIS).

Feature block	Macro-F1 with only this block	Δ macro-F1 when removed from the full set*
CSP	0.38	+0.015 (performance drops when removed)
Autocorrelation	0.42	−0.001
PSD	0.41	−0.003
Hjorth	0.40	−0.005

*Δ macro-F1 is defined as $F_1^{\text{all}} - F_1^{(-b)}$, where F_1^{all} is the macro-F1 with all blocks and $F_1^{(-b)}$ is the score with block b removed.

Several patterns emerge. First, when used in isolation, all four blocks produce macro-F1 scores in the range 0.38–0.42. Autocorrelation features are the strongest single source of discriminative information, slightly outperforming PSD and Hjorth statistics, which aligns with the intuition that imagined speech relies on intricate temporal dependencies rather than purely stationary band-power shifts. Second, the LOBO analysis reveals that CSP is the only block whose removal consistently harms the combined model: excluding CSP reduces macro-F1 by about 0.015, whereas excluding other blocks produces negligible or slightly positive changes at the third decimal place. In other words, CSP contributes a unique spatial structure that is not fully recoverable from the remaining blocks, even after dimensionality reduction, see Fig. 6.

These results suggest a division of labour across blocks. The temporal and spectral features (autocorrelation, PSD, and Hjorth) provide broadly redundant evidence about the underlying dynamics, which helps stabilise training but does not dramatically change the decision boundary when one block is removed. CSP, by contrast, contributes a smaller number of highly informative dimensions that appear to encode class-specific spatial patterns, consistent with its established role in motor imagery BCIs and more recent imagined-speech studies. From a design perspective, this analysis justifies keeping all four families in the reduced 3440-dimensional representation, but it also highlights that any future attempts at aggressive feature pruning should preserve CSP components and at least one of the temporal blocks to maintain robustness.

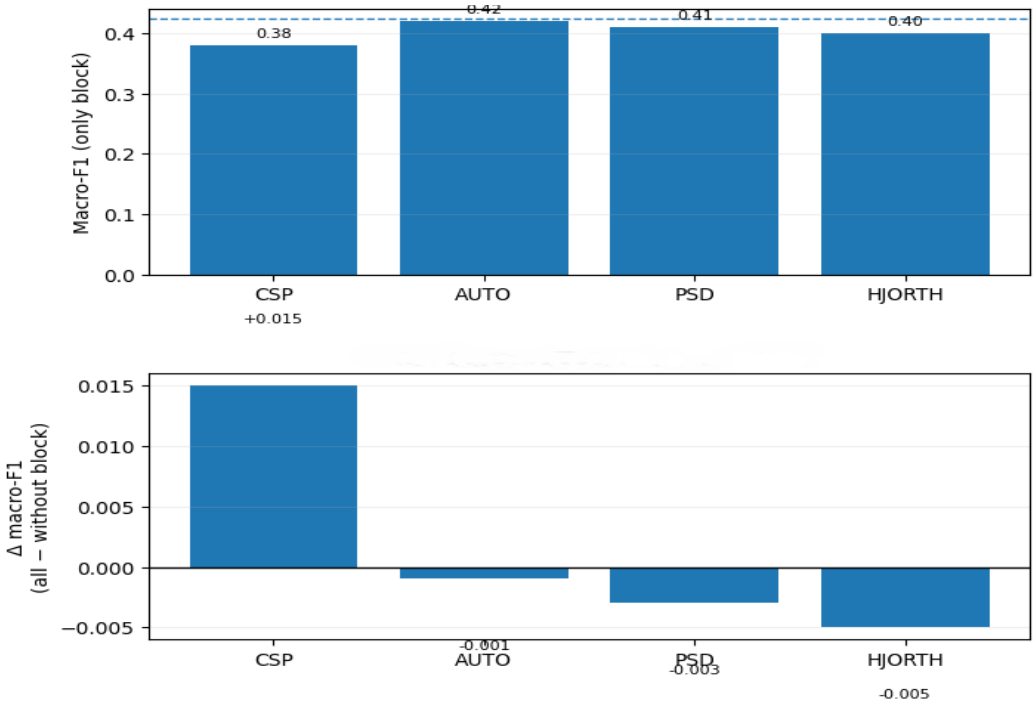


Fig. 6. Top: macro-F1 when each feature block (CSP, autocorrelation, PSD, Hjorth) is used in isolation. Bottom: change in macro-F1 when each block is removed from the full feature set.

C. Geometry of the Reduced Feature Space

To better understand why all three classifiers converge to similar performance levels, the geometry of the 3440-dimensional reduced feature space was examined. Fig. 7 visualises the test set using a two-dimensional t-SNE embedding that preserves local neighbourhood structure. Several loose clusters emerge, yet almost all of them are multi coloured. The three directional commands share overlapping territories, which reflects the symmetric confusions seen in the confusion matrix. The longer phrases occupy broader regions with slightly more homogeneous colour, consistent with their somewhat higher F1 scores, but there is no clear linear margin separating them from the short commands. The vowel trials disperse across multiple t-SNE islands with substantial intermixing, which again matches their poor separability at the classifier level.

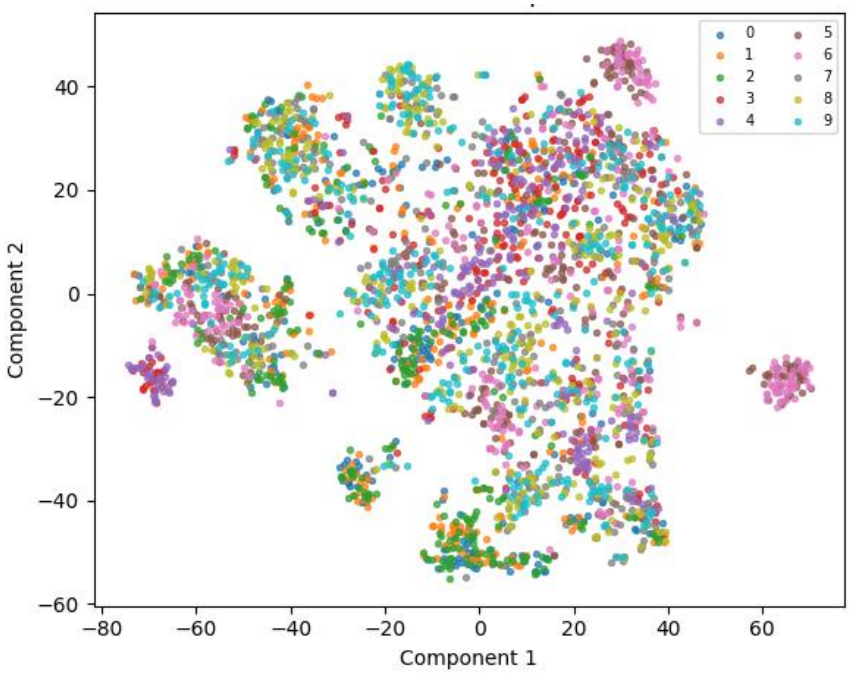


Fig. 7. t-SNE embedding of the reduced ASU feature space, coloured by class label, showing loosely clustered but strongly overlapping class manifolds.

The PCA projection in Fig. 8 provides a complementary, more linear view. The first principal component appears to capture a generic energy or amplitude factor: classes with higher overall band power, such as the composite phrases, are shifted toward the positive side of the axis, whereas short commands and vowels populate the centre and negative side. Along the second component the classes fan out in a wedge shaped pattern, but the degree of overlap remains high. This suggests that, although the block selection and dimensionality reduction pipeline has removed redundant and noisy dimensions, the remaining axes still encode a continuum of task difficulty rather than clearly separated class specific manifolds

Taken together, the t-SNE and PCA plots explain why the linear and non-linear models end up so close in performance. The reduced feature space is only weakly structured at the class level: there are pockets where neighbourhoods align with labels, particularly for some long phrases and composite commands, but large swathes of the space remain mixed. Similar overlapping manifolds have been reported in other imagined speech datasets, even when more aggressive representation learning or Riemannian geometry based features are used [36]. In that context, achieving around 0.41 macro F1 on a ten class, cross trial setting appears consistent with the intrinsic geometry of the data rather than a failure of the classifier family. This observation motivates our later shift toward transfer learning and column wise progressive architectures, where the goal is not only to learn a decision boundary in a fixed space but also to reshape that space through shared representations drawn from larger, more diverse corpora.

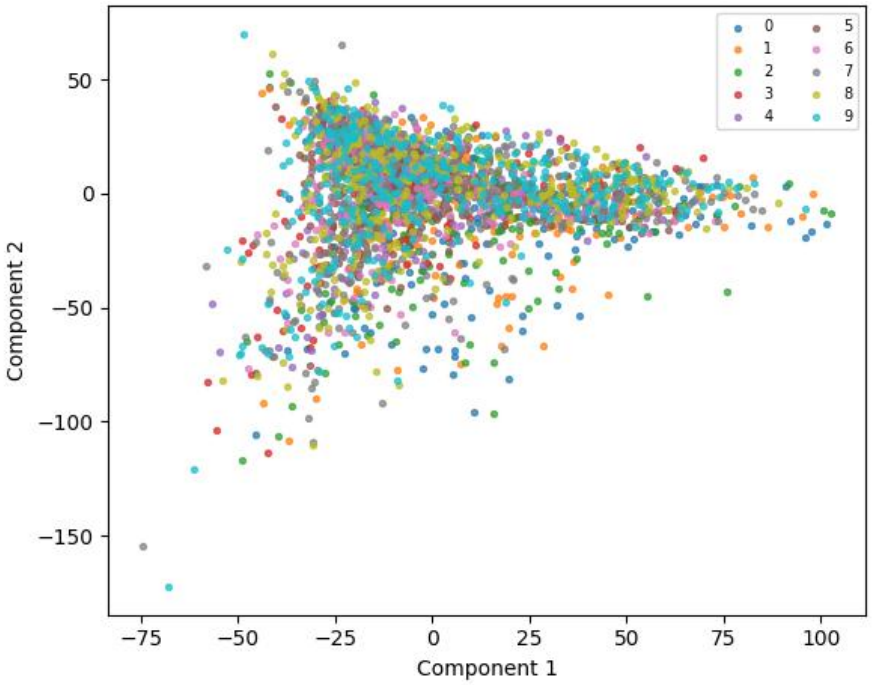


Fig. 8. First two principal components of the reduced ASU feature space, indicating a dominant energy dimension and persistent overla p between commands, phrases, and vowels.

D. Discussion

Taken together, these results confirm that, given the engineered feature space, model choice has limited impact: all three classifiers remain in a narrow band around 0.40–0.42 macro-F1. The ablations in Fig. 5 show that CSP contributes the only clearly non-redundant block, while temporal statistics are largely interchangeable, which explains why deeper models cannot extract much extra structure as described in Section IV-B. Within this context, the confusion matrix in Fig. 4 can be read as a stress test of the feature space. Sentences that differ in high-level semantics but share similar acoustic imagery, such as “in”, “out”, and “up”, are frequently confused, while commands that combine distinctive prosody and articulation, such as “cooperate-in”, are recognised more reliably. The vowels /a/, /i/, and /u/ form another tightly coupled group, with systematic pairwise confusions that align with their proximity in both articulatory and formant space. These patterns reinforce the message from the feature-block analysis: the current representation is sensitive to broad articulatory structure but struggles to separate fine-grained variations that are subtle even at the overt-speech level.

It is important to interpret the approximately 0.41 test accuracy and 0.42 macro-F1 in the light of how the ASU corpus has been used in earlier work. Most published studies restrict attention to a much smaller label space, typically binary contrasts such as “in” versus “cooperate” or three-class vowel sets, and often operate in subject-dependent mode with carefully selected trials. Under these conditions, reported accuracies frequently reach 70–90 per cent, but the task differs fundamentally from the ten-class, subject-wise split considered here. A cross-corpus studies, for example, demonstrates strong performance on ASU and related datasets, yet focuses on a reduced vocabulary and task structure that is closer to command-selection than to continuous imagined speech decoding [37]-[40].

Against this backdrop, the present results sit in the middle of what recent reviews describe as the "moderate-accuracy regime" for non-invasive imagined speech: performance is clearly above chance, robust across folds, and consistent between accuracy and macro-F1, yet still far from the reliability required for high-bandwidth communication [41]-[45] . Our contribution is less about pushing the absolute numbers and more about unpacking where the remaining errors come from. By combining block-wise ablations, classifier comparisons, learning curves, and low-dimensional visualisations, the figures in this section provide a multi-angle diagnosis that is often absent from earlier studies, where only a single classifier and headline accuracy are reported. This richer analysis shows that, for ASU, the bottleneck lies in the front-end representation and label design rather than in a lack of deep learning capacity [46]-[52].

Many imagined-speech systems that report high recognition rates do so within individual subjects, implicitly allowing models to specialise to idiosyncratic neural patterns. In contrast, the overlapping class manifolds observed here indicate that, once trials from multiple talkers and sessions are pooled, the neural signatures of individual words are less separable. This observation aligns with recent calls in the BCI literature to move away from over-optimistic subject-specific benchmarks and toward more realistic cross-subject evaluation protocols.

E. Limitations and Implications for Future Work

Several limitations follow directly from the figures and diagnostics presented above. First, all analyses are restricted to a single corpus. Although the ASU dataset is attractive because it combines ten imagined commands with controlled acquisition, it does not capture cross-linguistic variation, free-running internal speech, or very large vocabularies. The absence of cross-dataset validation means that we cannot yet claim that the identified feature blocks, such as autocorrelation and CSP, will remain optimal when transferred to other corpora like KARA ONE or BCI Competition datasets. Prior cross-corpus work suggests that substantial recalibration is usually required when switching between datasets with different channel layouts or cueing schemes.

Second, the present framework treats each trial as an isolated five-second segment and ignores finer temporal structure within the cue window. While this is appropriate for the feature blocks used here, it limits what the recurrent and convolutional layers can learn. Architectures that operate on lower-level time–frequency maps or directly on sensor-space waveforms might exploit temporal dynamics more fully, at the cost of more complex regularisation. Finally, the t-SNE and PCA plots are inherently descriptive. They suggest that class manifolds overlap and that some clusters are more compact than others, but they do not provide formal guarantees about separability or sample complexity.

Despite these constraints, the analysis points to several concrete directions. The redundancy revealed in Fig. 5 argues for more principled feature selection or sparsity-inducing penalties at the feature level, rather than only at the classifier weights. The confusion patterns in Fig. 4 suggest that future vocabulary design for imagined-speech BCIs should prioritise articulatorily distinct commands and avoid near-minimal pairs until more powerful representations are available. Finally, the modest yet stable performance of the logistic baseline indicates that well-regularised linear models remain a strong reference for imagined-speech EEG and should accompany any claims of deep-learning gains in future studies.

V. Conclusion

This work examined how far carefully engineered EEG features can support imagined speech recognition on the ASU corpus under a ten-class, cross-trial setting. Starting from 6,520 trials, a 3,440-dimensional representation was constructed that combines CSP spatial filters, temporal autocorrelation, PSD band power, and Hjorth parameters, and three classifiers were then evaluated on this representation: logistic regression, a three-layer MLP, and a CNN–BiLSTM–FCNN hybrid. All three models converged to a narrow performance band around 0.41 test accuracy and 0.42 macro-F1, suggesting that, within the evaluated protocol, the main bottleneck lies in the front-end representation and label structure rather than in classifier capacity. The block-wise ablations and geometric analyses clarify this constraint. CSP emerged as the only block whose removal clearly degraded performance on the full set, while temporal statistics were largely interchangeable, and both t-SNE and PCA revealed strongly overlapping class manifolds, especially among short commands and vowels. These patterns offer a plausible account of why deeper models failed to deliver systematic gains and place the present results within the "moderate-accuracy" regime reported for non-invasive imagined speech under realistic evaluation protocols. Several limitations temper the scope of these conclusions. First, the evaluation is confined to a single corpus (ASU); whether the block-level ranking, and in particular the primacy of CSP, holds under different channel layouts, cueing schemes and subject populations cannot be determined from this study alone. Second, all reported numbers are single-seed point estimates; confidence intervals from multi-seed and bootstrap resampling are not yet provided. Third, the deeper classifier treats a heterogeneous feature vector as a 1-D sequence, which is a conservative rather than an architecturally ideal choice for tabular inputs. Against these constraints, the study still provides a transparent and reproducible baseline, showing what can be achieved with well-understood EEG descriptors and conceptually simple classifiers, and pinpointing where progress is most needed. Several concrete next steps follow directly from the analysis. A first priority is cross-corpus validation, replicating the block-wise ablation on Kara One, FEIS and the BCI Competition imagined-speech track to test whether CSP remains the single non-redundant block under different montages and paradigms. A second priority is statistical robustness, repeating every model with at least five seeds and reporting bootstrap confidence intervals on accuracy and macro-F1. A third priority is the classifier side: gradient-boosted ensembles, regularised linear models with sparsity, and attention-based tabular architectures are a more principled match to the heterogeneous feature vector than the CNN–BiLSTM hybrid used here, and quantifying whether any of them escape the 0.42 macro-F1 ceiling on the same representation would be informative. A fourth priority is representation learning and transfer: progressive neural networks and similar architectures that reshape the feature space using larger and more diverse corpora remain the most plausible route to breaking out of the geometry described in Section IV D.

Finally, once a stable offline baseline is established across corpora, a real-time feasibility study – with streaming feature extraction and subject-adaptive calibration – would bridge the offline analysis presented here to the practical BCI setting. In this sense, the present analysis is intended as a diagnostic front end for subsequent stages of a broader programme toward reliable imagined-speech interfaces.

AUTHORS' CONTRIBUTION

The authors' contributions are as follows: Masrullizam Mat Ibrahim and Jamil Abedalrahim Jamil Alsayaydeh handled conceptualization; Safarudin Gazali Herawan was responsible for methodology; Jamil Abedalrahim Jamil Alsayaydeh and Hatem T M Duhair managed software development; Masrullizam Mat Ibrahim and Mazen Farid; validated and conducted formal analysis; Jamil Abedalrahim Jamil Alsayaydeh and Mazen Farid carried out the investigation; resources were provided by Hatem T M Duhair; the original draft was written by Jamil Abedalrahim Jamil Alsayaydeh and Hatem T M Duhair; Hatem T M Duhair reviewed and edited the writing; and Safarudin Gazali Herawan secured funding.

ACKNOWLEDGMENT

The authors express their gratitude to the Centre for Research and Innovation Management (CRIM) at Universiti Teknikal Malaysia Melaka (UTeM) for their valuable support in this research.

DATA AVAILABILITY STATEMENT

All the datasets used in this study are available from the Zenodo database (accession number: https://zenodo.org/records/17816093).

REFERENCES

[1] N. Fitriah, H. Zakaria, and T. L. E. Rajab, "EEG-Based Silent Speech Interface and its Challenges: A Survey," International Journal of Advanced Computer Science and Applications, vol. 13, no. 11, doi: 10.14569/IJACSA.2022.0131173.

[2] X.-Y. Liu et al., "Recent applications of EEG-based brain-computer-interface in the medical field," Military Medical Research, vol. 12, no. 1, p. 14, Mar. 2025, doi: 10.1186/s40779-025-00598-z.

[3] J. T. Panachakel and A. G. Ramakrishnan, "Decoding Covert Speech From EEG-A Comprehensive Review," Front. Neurosci., vol. 15, Apr. 2021, doi: 10.3389/fnins.2021.642251.

[4] K. Su and L. Tian, "Systematic review: progress in EEG-based speech imagery brain-computer interface decoding and encoding research," PeerJ Comput. Sci., vol. 11, p. e2938, Jun. 2025, doi: 10.7717/peerj-cs.2938.

[5] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features," J Neural Eng, vol. 15, no. 1, p. 016002, Feb. 2018, doi: 10.1088/1741-2552/aa8235.

[6] J. T. Panachakel, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "Decoding imagined speech using wavelet features and deep neural networks," in 2019 IEEE 16th India Council International Conference (INDICON), IEEE, 2019, pp. 1–4. Accessed: Dec. 05, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9028925/?casa_token=-IZJ8H3BhbUAAAAA:15xGxiarXLBBMqYwlHi-yAIccIACdvz8e_PX7-TqwwpQG293tZoBtqDPXCgKRGYRKlc5ySNgAg

[7] N. C. Mahapatra and P. Bhuyan, "Multiclass Classification of Imagined Speech Vowels and Words of Electroencephalography Signals Using Deep Learning," Advances in Human-Computer Interaction, vol. 2022, 2022, doi: 10.1155/2022/1374880.

[8] S. Alzahrani, H. Banjar, and R. Mirza, "Systematic Review of EEG-Based Imagined Speech Classification Methods," Sensors, vol. 24, no. 24, Art. no. 24, Jan. 2024, doi: 10.3390/s24248168.

[9] D. -Y. Lee, M. Lee, and S. -W. Lee, "Decoding Imagined Speech Based on Deep Metric Learning for Intuitive BCI Communication," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 1363–1374, 2021, doi: 10.1109/TNSRE.2021.3096874.

[10] M. M. Abdulghani, W. L. Walters, and K. H. Abed, "Imagined Speech Classification Using EEG and Deep Learning," Bioengineering (Basel), vol. 10, no. 6, p. 649, May 2023, doi: 10.3390/bioengineering10060649.

[11] Y. F. Alharbi and Y. A. Alotaibi, "Decoding Imagined Speech from EEG Data: A Hybrid Deep Learning Approach to Capturing Spatial and Temporal Features," Life, vol. 14, no. 11, Art. no. 11, Nov. 2024, doi: 10.3390/life14111501.

[12] M. Bisla and R. S. Anand, "Optimized CNN-Bi-LSTM–Based BCI System for Imagined Speech Recognition Using FOA-DWT," Advances in Human-Computer Interaction, vol. 2024, no. 1, p. 8742261, 2024, doi: 10.1155/2024/8742261.

[13] J. T. Panachakel and R. A. Ganesan, "Decoding Imagined Speech From EEG Using Transfer Learning," IEEE Access, vol. 9, pp. 135371–135383, 2021, doi: 10.1109/ACCESS.2021.3116196.

[14] A. Einizade, M. Mozafari, S. Jalilpour, S. Bagheri, and S. Hajipour Sardouie, "Neural decoding of imagined speech from EEG signals using the fusion of graph signal processing and graph learning techniques," Neuroscience Informatics, vol. 2, no. 3, p. 100091, Sept. 2022, doi: 10.1016/j.neuri.2022.100091.

[15] Z. Zhang et al., "Chisco: An EEG-based BCI dataset for decoding of imagined speech," Sci Data, vol. 11, no. 1, p. 1265, Nov. 2024, doi: 10.1038/s41597-024-04114-1.

[16] P. He, G. Wilson, and C. Russell, "Removal of ocular artifacts from electro-encephalogram by adaptive filtering," Med. Biol. Eng. Comput., vol. 42, no. 3, pp. 407–412, May 2004, doi: 10.1007/BF02344717.

[17] T. Memmott et al., "BciPy: brain–computer interface software in Python," Brain-Computer Interfaces, vol. 8, no. 4, pp. 137–153, Oct. 2021, doi: 10.1080/2326263X.2021.1878727.

[18] A. Akbarinia, Optimising EEG decoding with refined sampling and multimodal feature integration. 2024. doi: 10.48550/arXiv.2409.20086.

[19] G. Zhang, D. R. Garrett, and S. J. Luck, "Optimal filters for ERP research I: A general approach for selecting filter settings," Psychophysiology, vol. 61, no. 6, p. e14531, Jun. 2024, doi: 10.1111/psyp.14531

[20] M. Gyurkovics, G. M. Clements, K. A. Low, M. Fabiani, and G. Gratton, "The impact of 1/f activity and baseline correction on the results and interpretation of time-frequency analyses of EEG/MEG data: A cautionary tale," NeuroImage, vol. 237, p. 118192, Aug. 2021, doi: 10.1016/j.neuroimage.2021.118192.

[21] F. Taleb, M. Vasco, N. Rajabi, M. Björkman, and D. Kragic, "Challenging Deep Learning Methods for EEG Signal Denoising under Data Corruption," in 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), July 2024, pp. 1–4. doi: 10.1109/EMBC53108.2024.10782132.

[22] J. Yin, A. Liu, L. Wang, R. Qian, and X. Chen, "Integrating spatial and temporal features for enhanced artifact removal in multi-channel EEG recordings," J. Neural Eng., vol. 21, no. 5, p. 056018, Sept. 2024, doi: 10.1088/1741-2552/ad788d.

[23] B. Pester and C. Ligges, "Does independent component analysis influence EEG connectivity analyses?" in Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2018, pp. 1007–1010, doi: 10.1109/EMBC.2018.8512425.

[24] A. Echtioui, W. Zouch, M. Ghorbel, M. B. Slima, A. B. Hamida, and C. Mhiri, "Automated EEG Artifact Detection Using Independent Component Analysis," 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp. 1–5, Sept. 2020, doi: 10.1109/ATSIP49331.2020.9231574.

[25] X. Yin, M. Meng, Q. She, Y. Gao, and Z. Luo, "Optimal channel-based sparse time-frequency blocks common spatial pattern feature extraction method for motor imagery classification," Math. Biosci. Eng., vol. 18, no. 4, pp. 4247–4263, 2021, doi: 10.3934/mbe.2021213.

[26] B. Wang et al., "Common Spatial Pattern Reformulated for Regularizations in Brain-Computer Interfaces," IEEE Trans. Cybern., vol. 51, no. 10, pp. 5008–5020, 2021, doi: 10.1109/TCYB.2020.2982901.

[27] D. Pawar and S. Dhage, "Imagined Speech Classification using EEG based Brain-Computer Interface," in 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), IEEE, 2022, pp. 662–666. Accessed: Dec. 05, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9787644/?casa_token=q5YZSt-70HgAAAAA:3jTrKlxnQOJjSbaPu6dC9Ilr1B1hhSbia7GrKko6EdG9lSa-blnvzPpn7EBaWnD0JjEkF1fd7A

[28] A. Moura, S. Lopez, I. Obeid, and J. Picone, "A comparison of feature extraction methods for EEG signals," in 2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Dec. 2015, pp. 1–2. doi: 10.1109/SPMB.2015.7405430.

[29] I. Carrara and T. Papadopoulo, "Classification of BCI-EEG based on augmented covariance matrix," 2023, doi: 10.48550/ARXIV.2302.04508.

[30] S.-H. Lee, M. Lee, and S.-W. Lee, "Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 12, pp. 2647–2659, 2020.

[31] W. H. Alawee, A. Basem, and L. A. Al-Haddad, "Advancing biomedical engineering: Leveraging Hjorth features for electroencephalography signal analysis," J Electr Bioimpedance, vol. 14, no. 1, pp. 66–72, doi: 10.2478/joeb-2023-0009.

[32] O. George et al., "State-of-the-Art Versus Deep Learning: A Comparative Study of Motor Imagery Decoding Techniques," IEEE Access, vol. 10, pp. 45605–45619, 2022, doi: 10.1109/ACCESS.2022.3165197.

[33] F. Qi, W. Wu, K. Liu, T. Yu, and Y. Cao, "A Logistic Regression Based Framework for Spatio-Temporal Feature Representation and Classification of Single-Trial EEG," May 2021. doi: 10.1007/978-981-16-2336-3_36.

[34] D. Alonso-Vázquez, O. Mendoza-Montoya, R. Caraza, H. R. Martinez, and J. M. Antelis, "From pronounced to imagined: improving speech decoding with multi-condition EEG data," Front. Neuroinform., vol. 19, June 2025, doi: 10.3389/fninf.2025.1583428.

[35] P. Saha, S. Fels, and M. Abdul-Mageed, "Deep Learning the EEG Manifold for Phonological Categorization from Active Thoughts," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, pp. 2762–2766. doi: 10.1109/ICASSP.2019.8682330.

[36] M. A. Bakhshali, M. Khademi, A. Ebrahimi-Moghadam, and S. Moghimi, "EEG signal classification of imagined speech based on Riemannian distance of correntropy spectral density," Biomedical Signal Processing and Control, vol. 59, 2020, doi: 10.1016/j.bspc.2020.101899.

[37] J. T. Panachakel, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "A novel deep learning architecture for decoding imagined speech from EEG," arXiv:2003.09374, Mar. 2020.

[38] H. T. M. Duhair, M. bin mat Ibrahim, J. A. J. Alsayaydeh, R. Bacarra, and A. H. Ahmed, "Progressive transfer learning Unifies Multi-Corpus EEG for robust and scalable Imagined-Speech decoding," Biomedical Signal Processing and Control, vol. 116, p. 109475, May 2026, doi: 10.1016/j.bspc.2026.109475.

[39] J. A. J. Alsayaydeh, W. A. Y. Khang, W. A. Indra, V. Shkarupylo and J. Jayasundar. 2019. Development of smart dustbin by using apps. ARPN Journal of Engineering and Applied Sciences. 14(21): 37033711.

[40] A. A. AlZubi, "Bunch graph based dimensionality reduction using auto-encoder for character recognition," Multimedia Tools and Applications, 2022.

[41] F. Lotte et al., "A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update," J Neural Eng, vol. 15, no. 3, p. 031005, June 2018, doi: 10.1088/1741-2552/aab2f2.

[42] J. A. J. Alsayaydeh, W. A. Indra, A. W. Y. Khang, V. Shkarupylo, and D. A. P. P. Jkatisan, "Development of vehicle ignition using fingerprint," ARPN Journal of Engineering and Applied Sciences, vol. 14, no. 23, pp. 4045–4053, 2019. [Online]. Available: http://www.arpnjournals.org/jeas/research_papers/rp_2019/jeas_1219_8024.pdf

[43] L. Kovalchuk, D. Kaidalov, A. Nastenko, M. Rodinko, O. Shevtsov, and R. Oliynykov, "Decreasing security threshold against double spend attack in networks with slow synchronization," Computer Communications, vol. 154, pp. 75–81, 2020, doi: 10.1016/j.comcom.2020.01.079.

[44] J. A. J. Alsayaydeh, W. A. Indra, A. W. Y. Khang, A. K. M. Z. Hossain, V. Shkarupylo, and J. Pusppanathan, "The experimental studies of the automatic control methods of magnetic separators performance by magnetic product," ARPN Journal of Engineering and Applied Sciences, vol. 15, no. 7, pp. 922-927, 2020.

[45] N. F. B. A. Rahim, A. W. Y. Khang, A. Hassan, S. J. Elias, J. A. M. Gani, J. Jasmis, and J. A. J. Alsayaydeh, "Channel Congestion Control in VANET for Safety and Non-Safety Communication: A Review," in 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2021, pp. 1-6, doi: 10.1109/ICRAIE52900.2021.9704017.

[46] J. A. J. Alsayaydeh et al., "Development of vehicle door security using smart tag and fingerprint system," International Journal of Engineering and Advanced Technology, vol. 9, no. 1, pp. 3108–3114, Oct. 2019, doi: 10.35940/ijeat.E7468.109119

[47] A. Mehmood, A. Abugabah, A. A. AlZubi, and L. Sanzogni, "Early diagnosis of Alzheimer's disease based on convolutional neural networks," Computer Systems Science and Engineering, vol. 43, no. 1, pp. 305–315, 2022.

[48] J. A. Alsayaydeh, M. Nj, S. N. Syed, A. W. Yoon, W. A. Indra, V. Shkarupylo and C. Pellipus, "Homes appliances control using bluetooth," ARPN Journal of Engineering and Applied Sciences, vol. 14 (19), pp. 3344-3357, 2019.

[49] A. Al Smadi, S. Yang, A. Abugabah, A. A. AlZubi, and L. Sanzogni, "A pansharpening based on the non-subsampled contourlet transform and convolutional autoencoder: Application to QuickBird imagery," IEEE Access, vol. 10, pp. 44778–44788, 2022.

[50] V. Shkarupylo, I. Blinov, A. Chemeris, V. Dusheba, J. A. J. Alsayaydeh, and A. Oliinyk, "Iterative approach to TLC model checker application," in 2021 IEEE 2nd KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2021, pp. 283–287, doi: 10.1109/KhPIWeek53812.2021.9570055.

[51] H. T. M. Duhair, M. Bin Mat Ibrahim, M. Farid, J. A. J. Alsayaydeh, and S. G. Herawan, "EEG-Based Imagined-Speech Decoding: A Review," Intl. J. Adv. Comput. Sci. Appl., vol. 17, no. 1, pp. 380–390, 2026, doi: 10.14569/IJACSA.2026.0170136.

[52] A. Karatkevich and I. Grobelna, "Deadlock detection in Petri nets: One trace for one deadlock?," 2014 7th International Conference on Human System Interactions (HSI), Costa da Caparica, Portugal, 2014, pp. 227-231, doi: 10.1109/HSI.2014.6860480.

A Multivocal Literature Review of Metadata Governance: Conceptual Foundations and Research Gaps

Dana Indra Sensuse[1], Alivia Yulfitri[2], Erisva Hakiki Purwaningsih[3], Anton Satria Prabuwono[4]
Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia[1,2,3]
Faculty of Computer Science, Universitas Esa Unggul, Jakarta, Indonesia[2]
Agency for Human Resources Development on Communications and Digital Affairs, Ministry of On comm and digital of the Republic of Indonesia, Jakarta, Indonesia[3]
Faculty of Science, Management & Computing, Universiti Teknologi PETRONAS, Perak Darul Ridzuan, Malaysia[4]

***Abstract*—As digital ecosystems grow and cross-organizational data sharing intensifies, metadata governance has become increasingly important. It plays a critical role in supporting interoperability, accountability, and sustainable data ecosystems. However, metadata governance remains conceptually fragmented and insufficiently structured as a distinct research domain. This study conducts a multivocal literature review (MLR) that integrates academic publications with authoritative non-academic sources, including international standards, governance frameworks, and regulatory instruments. A systematic review process was applied to identify governance constructs, recurring patterns, and conceptual gaps across heterogeneous evidence sources. The findings show that no peer-reviewed studies explicitly integrate metadata governance with formal conceptual structuring. This finding should not be interpreted as a limitation of the search process, but rather as an empirical indication of a structural gap in the literature. It provides a methodological justification for extending the analysis to multivocal evidence. The results further indicate that existing studies predominantly emphasize metadata management and technical interoperability. In contrast, governance-level constructs—such as decision rights, accountability mechanisms, oversight structures, and lifecycle coordination—remain underdefined and inconsistently formalized. This study synthesizes fragmented knowledge into a coherent conceptual understanding of metadata governance, clarifies its distinction from metadata management, and identifies critical research gaps. These findings provide a structured foundation for advancing metadata governance as a cumulative research domain and support future conceptual and methodological research.**

***Keywords*—Metadata governance; multivocal literature review; data governance; metadata management; conceptual synthesis**

I. Introduction

As digital ecosystems expand and cross-organizational data exchange intensifies, governance-related issues concerning metadata have become increasingly important. However, metadata governance remains insufficiently established as an explicit and formally articulated research domain.

Metadata plays a foundational role in enabling semantic interoperability, as discussed by Foulonneau and Riley [1] and further emphasized in subsequent studies on metadata practices [2], [3]. With the acceleration of digital transformation and platform-based ecosystems, metadata has evolved beyond technical documentation into a strategic asset influencing reliability, regulatory compliance, and long-term sustainability of data infrastructures [4], [5].

Despite its strategic importance, metadata practices remain predominantly governed by schemas, registries, catalogs, and interoperability standards. Governance concerns - such as decision authority, accountability allocation, policy enforcement, risk management, and lifecycle oversight - remain weakly articulated, fragmented across technical practices, or embedded within implementation-specific contexts. Widely adopted interoperability and registry standards emphasize syntactic and structural alignment rather than formal governance architectures [6], [7], [8], [9], [10], [11], while regulatory instruments on security and personal data protection impose compliance obligations without establishing explicit governance models for metadata [12], [13], [14], [15].

From an academic perspective, metadata research largely concentrates on metadata management, interoperability engineering, repository architectures, and domain-specific applications, whereas governance dimensions remain under-theorized and inconsistently implemented [1], [3]. Practitioner frameworks and empirical studies typically provide operational checklists, maturity models, and tooling guidance but rarely define governance structures that clarify decision rights, accountability mechanisms, and control models [16], [17]. Consequently, metadata governance has not yet emerged as a consolidated scientific domain supported by shared conceptual frameworks or validated structural models.

At the policy and infrastructure level, international organizations consistently identify inconsistent metadata standards and weak cross-organizational coordination as major barriers to interoperability and data reuse in digital government and open data ecosystems [18], [19], [20]. Similar governance challenges are evident in both national and global contexts. Empirical studies in Indonesia report persistent difficulties in metadata harmonization, interoperability, and institutional coordination despite regulatory mandates [21], [22], while international assessments highlight metadata governance capability as a critical determinant of interoperable data ecosystems [23].

Conceptually, the absence of integrated frameworks and formal metamodels further exacerbates fragmentation in metadata governance research and practice. To clarify the analytical scope of this study, conceptual structuring refers to the explicit representation of governance constructs and their relationships in a systematic and formalized manner.

Although conceptual frameworks and metamodels are widely used in enterprise architecture and information systems to structure domains and enable traceability [24], their systematic application in metadata governance remains limited. Existing standards and professional guidance address isolated components - such as registries, schemas, stewardship roles, or compliance controls - without consolidating them into coherent governance architectures or formalized conceptual models. This fragmentation hinders validation, reuse, cross-context comparison, and cumulative theoretical development.

Preliminary systematic searches indicate a significant lack of peer-reviewed studies explicitly integrating metadata governance with formal modeling or metamodeling approaches. This gap motivates the inclusion of authoritative grey literature capturing standards, industry frameworks, and regulatory instruments [25], and justifies the use of a multivocal literature review (MLR) to synthesize academic and practice-driven knowledge. This is particularly relevant because governance constructs are often institutionalized in standards, regulatory frameworks, and professional practices rather than in peer-reviewed academic publications.

Despite the increasing recognition of metadata governance in both academic and practitioner domains, there is still a lack of measurable and formally defined structures that characterize its maturity as a research domain. Existing studies do not provide 1) formally defined governance structures, 2) explicit decision-right models, and 3) clearly articulated conceptual representations of metadata governance.

This limitation is empirically supported by the systematic review results presented in Section IV, where zero eligible peer-reviewed studies were identified that explicitly integrate metadata governance with formal conceptual modeling or metamodeling approaches. The absence of formal conceptualization, combined with fragmented terminology and inconsistent articulation of governance constructs across sources, suggests limited theoretical consolidation rather than mere publication scarcity. This absence indicates not only a research gap, but also a lack of conceptual consolidation and methodological rigor within the field. This outcome was consistent across multiple query formulations and databases, indicating robustness of the search process rather than query bias.

Accordingly, this study conducts a multivocal literature review integrating scholarly publications and grey literature to examine the conceptual foundations of metadata governance. The study is guided by the following research questions:

RQ1. What thematic areas and research focal points have been explored in studies related to metadata governance and associated metamodeling perspectives?

RQ2. How thoroughly do current studies address the governance aspects of metadata, encompassing domain scope, process structures, and conceptual components?

RQ3. To what extent are validation approaches reported in studies related to metadata governance and its conceptual representation?

RQ4. What conceptual and methodological gaps remain in advancing metadata governance as a coherent research domain?

The remainder of this paper is structured as follows. Section II presents the conceptual background. Section III describes the research methodology. Section IV reports the findings and synthesis. Section V discusses implications and future directions, followed by the conclusion in Section VI.

This study contributes by synthesizing dispersed governance evidence through a multivocal literature review, clarifying the conceptual boundaries between metadata governance and metadata management, and providing a structured conceptual foundation for future research on metadata governance.

II. Conceptual Background

A. Governance and Management

Governance is commonly defined as the system through which organizations are directed and controlled, encompassing the allocation of decision authority, accountability structures, and oversight mechanisms that guide strategic outcomes. The G20/OECD Principles of Corporate Governance frame governance as the institutional arrangement through which objectives are set, performance is monitored, and accountability relationships are structured among governing bodies, management, and stakeholders [26].

A fundamental distinction in governance literature separates governance from management. Governance focuses on setting direction, allocating decision rights, and ensuring oversight, whereas management concerns the planning and execution of operational activities within that direction [27]. In the data domain, Khatri and Brown conceptualize governance as the allocation of decision rights and accountabilities across key decision areas, emphasizing who has authority to decide and who is responsible for outcomes [28].

This distinction is further operationalized in IT governance standards. ISO/IEC 38500 defines governance as the system by which organizational use of technology is directed and controlled, positioning governance as a leadership function responsible for evaluating strategic options, directing management, and monitoring outcomes [29]. Similarly, COBIT distinguishes governance from management through the Evaluate–Direct–Monitor (EDM) domain, which represents the responsibilities of governing bodies in evaluating strategic needs, directing priorities, and monitoring performance and compliance [30], [31].

Together, these perspectives establish governance as a hierarchical and oversight-oriented function concerned with decision authority, accountability, and monitoring.

This governance–management distinction provides the conceptual foundation for positioning metadata governance as an institutional function focused on decision rights and accountability rather than purely operational metadata management.

B. Metadata and Metadata Management

Metadata is commonly defined as structured information that describes, explains, locates, or otherwise facilitates the management, use, and understanding of data resources [1], [7]. It provides semantic context, provenance information, quality indicators, and structural descriptions that enable interoperability and meaningful reuse across heterogeneous systems [2], [3].

Within the data management discipline, DAMA-DMBOK defines metadata as information that describes data assets and provides context for their meaning, quality, structure, ownership, lineage, and usage across the organization [32]. DAMA emphasizes that metadata represents a critical organizational asset because it enables consistency, discoverability, governance enforcement, and integration across enterprise data ecosystems.

DAMA further defines metadata management as the discipline responsible for planning, implementing, and controlling the activities required to create, maintain, integrate, and govern metadata throughout its lifecycle [32]. Metadata management encompasses activities such as metadata modeling, standards definition, registry and catalog management, lineage tracking, quality profiling, and metadata integration with operational and analytical platforms.

International standards complement this perspective by formalizing metadata structures and interoperability mechanisms. ISO/IEC 11179 specifies the structure and management of metadata registries, while ISO/IEC 19763 addresses interoperability and model-driven metadata integration across heterogeneous systems [7], [10]. These standards primarily focus on syntactic consistency, semantic alignment, and technical interoperability rather than institutional governance mechanisms.

Although metadata management establishes the operational foundation for managing metadata assets, it does not explicitly address governance questions such as decision authority, accountability allocation, policy enforcement, compliance assurance, risk oversight, and long-term institutional sustainability. As a result, organizations may achieve technically mature metadata infrastructures while lacking coherent governance structures that ensure strategic alignment, accountability, and cross-organizational consistency.

C. Data Governance and its Relationship to Metadata

Building on the governance foundations discussed earlier, data governance is commonly conceptualized as the institutional arrangement of decision rights, accountabilities, policies, and control mechanisms that ensure data is managed as a strategic organizational asset. From a design perspective, Khatri and Brown conceptualize data governance through explicit allocation of decision rights across key decision domains, including data principles, data quality, access and security, regulatory compliance, and data architecture [28]. These decision domains are operationalized through decision-rights matrices to clarify authority and accountability within organizations.

Complementing decision-rights perspectives, governance has also been conceptualized as an organizational capability that aligns roles, responsibilities, and governance processes to sustain data quality and regulatory compliance [33]. Accountability structures are often operationalized through responsibility assignment mechanisms, such as RACI models, which clarify authority and auditability across governance domains [34]. Practitioner-oriented frameworks further consolidate governance design into integrated operating models encompassing organizational structures, policies, processes, and performance metrics [35].

Within enterprise data management bodies of knowledge, DAMA defines data governance as the function that establishes accountability, policies, and decision rights to ensure effective data management across the organization [32]. In DAMA's framework, metadata is explicitly governed under the scope of data governance but operationalized through the metadata management knowledge area, highlighting policy enforcement, stewardship roles, and governance structures as core mechanisms.

Although metadata is recognized as a critical enabling component supporting discoverability, traceability, and auditability, governance frameworks generally treat metadata as a supporting artifact rather than as a governance object in its own right. Decision structures and control mechanisms are primarily defined at the data asset level without formalizing governance structures specifically for metadata. This implicit positioning creates conceptual ambiguity: while data governance establishes general governance principles, it does not sufficiently articulate how metadata itself should be governed as an institutional asset with its own lifecycle, risks, and compliance implications.

Empirical and practitioner studies indicate that weaknesses in metadata management and governance directly affect decision quality, operational risk, and compliance exposure [4], [5]. Yet, these studies rarely articulate explicit governance structures for metadata, reinforcing the need for a dedicated conceptual consolidation of metadata governance.

D. Metadata Governance as an Emerging Domain

Building on the governance foundations discussed above, metadata governance can be understood as the system of policies, roles, decision rights, controls, and assurance mechanisms that regulate how metadata is defined, maintained, shared, protected, and evolved across organizational boundaries. This definition represents a synthesized conceptual interpretation derived from the literature, rather than an established or universally accepted definition. Unlike metadata management, which focuses on operational execution, metadata governance addresses institutional authority, accountability, compliance, and long-term sustainability. In this study, metadata governance is positioned not merely as an extension of data governance applied to metadata, but as a distinct governance object with its own decision scope, lifecycle considerations, and risk profile.

Explicit use of the term metadata governance remains limited in both academic and practitioner sources. Among established bodies of knowledge and professional frameworks, explicit conceptualization of metadata governance as a distinct domain remains largely absent. DAMA-DMBOK is currently the only widely recognized reference that explicitly uses the term, yet even within this framework, metadata governance is not treated as an independent domain but is embedded within metadata management [32]. This reflects a limitation in the current body of knowledge, where metadata governance has not yet been formalized as an independent conceptual domain across multiple authoritative sources.

In contrast, broader governance frameworks such as ISO/IEC 38500 and COBIT 2019 provide general governance principles - such as decision rights, accountability, and oversight - but do not explicitly address metadata governance as a distinct conceptual domain. This indicates a structural gap in the literature rather than selective reliance on a single reference.

In academic literature, governance dimensions are typically addressed only implicitly and are not formalized as explicit governance constructs or models. Metadata-related studies typically focus on interoperability, quality, registries, or domain-specific standards rather than governance structures [1], [3]. Practitioner frameworks and industry sources provide operational guidance on stewardship and tooling, but seldom articulate governance mechanisms in a systematic or formalized manner [16], [17].

Regulatory and standardization documents [12], [13] introduce obligations related to security, privacy, and accountability that indirectly affect metadata governance. However, these instruments do not provide integrated governance architectures or conceptual models specific to metadata, reinforcing the absence of a formalized governance structure.

Taken together, existing sources predominantly frame metadata governance as fragmented operational practices and tool-oriented guidance rather than as a governance-level construct grounded in explicit decision rights, accountability structures, and institutional control mechanisms. This fragmentation indicates that metadata governance remains an emergent domain without a consolidated theoretical or conceptual foundation. As a result, metadata governance has not yet consolidated into a coherent research domain with shared terminology, validated models, or cumulative theoretical development.

III. Research Methodology

A. Design of the Study

This study adopts a Multivocal Literature Review (MLR) design with qualitative document analysis to examine the state of the art of metadata governance as an emerging scientific domain [36], [37], [38]. The MLR approach is appropriate because relevant knowledge is distributed across both academic publications and non-academic sources.

The analysis focuses on governance-level constructs, including decision rights, accountability mechanisms, institutional roles, policy instruments, and formal control structures, while excluding purely technical implementation aspects.

The review process consists of four stages. First, relevant sources are systematically identified and screened using explicit inclusion and exclusion criteria. Second, governance-related constructs are extracted from eligible documents using qualitative document analysis. Third, extracted constructs are synthesized to identify recurring patterns, conceptual convergence and divergence, and structural gaps across academic literature, grey literature, and authoritative non-academic sources such as standards, regulatory documents, and governance frameworks. Finally, the synthesis is interpreted to produce a state-of-the-art mapping and to derive implications and a future research agenda. These stages are further operationalized in the subsequent subsections (Sections C–F), while Section B describes the overall review protocol and its methodological rationale.

B. Review Protocol and Rationale

The review protocol follows established guidance for systematic literature reviews [25] and multivocal literature reviews to ensure transparency and rigor [36], [37].

In this study, MLR integrates academic literature identified through database search with non-academic sources. Grey literature refers to non-peer-reviewed practitioner-oriented materials, while standards, regulatory documents, and governance frameworks are treated as authoritative sources providing normative perspectives.

The protocol operationalizes the four review stages described in Section A (identification, screening, extraction, and synthesis), which are further detailed in the subsequent subsections (Sections C–F).

C. Search Strategy and Data Sources

A two-stage search strategy was applied. Academic sources were retrieved from Scopus, IEEE Xplore, ScienceDirect, SpringerLink, SAGE Journals, and Taylor & Francis, covering publications from 2000 to 2025.

Non-academic sources were identified through targeted search using general search engines, Google Scholar, and official repositories of standards bodies, professional organizations, and regulatory institutions. This approach ensures coverage of heterogeneous evidence beyond peer-reviewed publications and is consistent with multivocal literature review practices [36], [37].

D. Inclusion and Exclusion Criteria

Sources were selected based on governance-level relevance and conceptual structuring of metadata. Academic sources were included if they were peer-reviewed publications (journal or conference papers), published in English, and explicitly addressed metadata governance and/or conceptual modeling perspectives, including structural or metamodel-related representations. Studies focusing solely on data governance or operational metadata management without governance-level abstraction were excluded.

Non-academic sources were included when originating from authoritative and traceable institutions, including grey literature,

standards bodies, professional organizations, and regulatory agencies. Sources lacking institutional authority or focusing purely on vendor-specific tooling or technical implementation were excluded to maintain a governance-level analytical focus.

E. Screening and Selection Process

Screening followed staged relevance filtering, including title and abstract screening followed by full-text assessment [25]. Duplicate records were removed during the screening process to ensure source uniqueness.

F. Data Extraction and Synthesis

Data extraction was conducted using qualitative document analysis to identify governance-related constructs, including decision rights, accountability structures, institutional roles, policy instruments, and control mechanisms [38]. Extracted constructs were iteratively compared and synthesized to identify recurring patterns, conceptual convergence and divergence, and structural gaps across academic literature, grey literature, and authoritative non-academic sources. Particular attention was given to distinguishing governance-level constructs from operational metadata management practices.

G. Quality Assessment and Source Evaluation

Quality assessment ensures source credibility, while analytical selection ensures conceptual relevance. Each selected source was evaluated based on three criteria: governance relevance, conceptual rigor, and evidence transparency. Governance relevance refers to the degree to which governance constructs are explicitly addressed. Conceptual rigor refers to the presence of structured models, frameworks, or formal definitions. Evidence transparency refers to the clarity of methodology, traceability of claims, and credibility of the source.

The quality assessment was conducted by the authors based on predefined criteria and was not intended as an objective measurement but as a structured analytical tool. The scoring supports consistent comparison across heterogeneous sources and reduces potential bias in interpreting multivocal evidence.

Each criterion was assessed using a three-level scale (low = 1, medium = 2, high = 3). To ensure consistency, each level was operationalized using indicative criteria. A high rating indicates that the source explicitly defines governance constructs, presents structured models or frameworks, and provides traceable and well-documented evidence. A medium rating indicates partial or implicit treatment of governance constructs, moderate conceptual structuring, or limited methodological transparency. A low rating indicates minimal or indirect reference to governance concepts, lack of formal structuring, or insufficient evidence transparency.

The final score reflects the overall contribution of the source to governance-level analysis. Lower-scoring sources were not excluded automatically, but were used cautiously to complement higher-quality sources, particularly when representing current industry practices. This approach ensures that the study's conclusions are not disproportionately influenced by weaker or purely descriptive sources.

Analytical roles were assigned independently from quality scores. Core sources were those that directly discuss metadata governance or metadata governance-related practices. Supporting sources were used to provide governance, compliance, security, or regulatory foundations for interpreting metadata governance. Contextual sources were used to explain metadata structures or technical foundations but were not used as primary governance evidence.

This quality assessment approach is adapted from established systematic literature review practices [25] and adjusted for multivocal literature review contexts [36], [37], where heterogeneous evidence sources must be evaluated consistently.

H. Analytical Selection Criteria

Following the quality assessment, a second-stage selection was applied to identify sources for core conceptual analysis. This stage focuses on governance-level relevance, particularly whether a source explicitly employs the term "metadata governance" or presents governance constructs such as decision rights, accountability structures, and control mechanisms.

Sources that primarily address operational, technical, or capability aspects without governance-level articulation were not included in the core analysis, even if they achieved high quality scores.

IV. Result

A. Results of the Systematic Literature Review (SLR)

The systematic literature review (SLR) was conducted across six academic databases using three query formulations targeting metadata governance and conceptual structuring. Table I reports the search outcomes by query and database, followed by title/abstract screening and full-text review. As shown in Table I, a total of 89 records were retrieved.

However, none of the retrieved studies were found to meet the inclusion criteria after title/abstract screening and full-text review. This result indicates that no peer-reviewed studies explicitly integrate metadata governance with formal conceptual structuring or metamodeling perspectives. This consistency across databases and query variations strengthens the validity of the result and reduces the likelihood of search bias or query formulation limitations.

This outcome should not be interpreted solely as a negative result, but rather as an indication of a conceptual and terminological gap in the literature. It reflects a conceptual gap in the literature, where 1) the term "metadata governance" is inconsistently used, 2) governance is often subsumed under broader data governance, and 3) governance perspectives are not integrated with formal modeling approaches. Therefore, this result provides a methodological justification for extending the evidence base beyond academic sources.

B. Multivocal Corpus and Evidence Landscape

Given the absence of eligible academic studies, the analysis was extended to non-academic and authoritative sources using a multivocal literature review approach.

TABLE I. SLR RESULTS BY QUERY AND DATABASE

No.	Query Description	Database	Initial Hits	After Title/Abstract Screening	After Full-Text Review
1	Metadata governance + metamodel terms	Scopus	0	0	0
		Taylor & Francis	0	0	0
		SAGE	0	0	0
		IEEE Xplore	0	0	0
		SpringerLink	0	0	0
		ScienceDIrect	0	0	0
2	Metadata governance + modeling frameworks	Scopus	16	0	0
		Taylor & Francis	9	0	0
		SAGE	6	0	0
		IEEE Xplore	0	0	0
		SpringerLink	27	0	0
		ScienceDirect	18	0	0
3	Metadata governance + conceptual structure	Scopus	7	0	0
		Taylor & Francis	1	0	0
		SAGE	0	0	0
		IEEE Xplore	0	0	0
		SpringerLink	1	0	0
		ScienceDirect	4	0	0
	TOTAL		89	0	0

These sources collectively represent heterogeneous evidence spanning structural, institutional, and practice-oriented perspectives. Standards (e.g., ISO/IEC 11179, ISO/IEC 19763) provide structural foundations for metadata representation, while governance frameworks (e.g., DAMA-DMBOK2, COBIT, ISO/IEC 38500, DCAM [39]) offer conceptual grounding for governance roles, controls, and accountability. Regulatory sources (e.g., GDPR, national data regulations) contribute compliance perspectives, and industry/practitioner sources reflect current implementation practices.

Table II presents the consolidated corpus of sources included in the multivocal literature review. The corpus reflects a heterogeneous evidence base, including academic literature, grey literature, international standards, governance frameworks, and regulatory sources.

TABLE II. CONSOLIDATED CORPUS OF SOURCES IN THE MULTIVOCAL LITERATURE REVIEW

No	Source	Type	Source Category	Rationale for Inclusion
1	DAMA-DMBOK2 Revised (2024) [32]	Industry Framework	Governance Framework	Primary reference for metadata management and governance practices
2	ISO/IEC 38500: Governance of IT for the Organization [29]	International Standard	Governance Standard	Provides governance principles (EDM, accountability)
3	COBIT 2019 [30], [31]	Industry Framework	Governance Framework	Defines governance domains and control structures
4	ISO/IEC 11179: Metadata Registries [7], [8], [9], [11]	International Standard	Metadata Standard	Defines metadata registry structures
5	ISO/IEC 19763: Metamodel Framework for Interoperability [10]	International Standard	Metadata Standard	Defines metamodel for interoperability
6	ISO 19115: Metadata for Geographic Information [6]	International Standard	Metadata Standard	Provides domain-specific metadata structure
7	ISO/IEC 27001: Information Security Management [13]	International Standard	Security Standard	Provides security governance context
8	General Data Protection Regulation (GDPR (EU 2016)) [12]	Regulation	Regulatory Source	Defines compliance and accountability obligations
9	Presidential Regulation of Indonesia No. 39/2019 on One Data Indonesia (Perpres 39/2019) [14]	Regulation	Regulatory Source	National data governance regulation
10	Law No. 27/2022 on Personal Data Protection (PDP Law) [15]	Regulation	Regulatory Source	Personal data protection framework
11	Metadata Governance in Media Industry [16]	Industry Study	Practice-Based Evidence	Demonstrates governance practices in industry
12	Atlan [40]	Industry Source	Grey Literature	Vendor-based governance practices
13	data.world [41]	Industry Source	Grey Literature	Metadata governance tooling and practices
14	Decube [42]	Industry Source	Grey Literature	Governance implementation practices
15	Dataversity [43]	Industry Source	Grey Literature	Educational and practitioner materials

16	DCAM v2.2 [39]	Industry Framework	Governance Framework	Data governance capability model
17	MetaStore [17]	Industry Framework	Academic	Metadata architecture model
18	COBIT-based IT Governance Metamodel [44]	Academic Model	Academic	Formal governance metamodel (reference)

In addition, practitioner-oriented sources such as Atlan [40], data.world [41], Decube [42], and Dataversity [43] illustrate how the term "metadata governance" is used in practice, often referring to operational and tool-oriented activities rather than formal governance structures.

C. Quality Assessment and Analytical Selection

Following corpus construction, all sources were evaluated using a structured quality assessment framework (Table III). The assessment considers governance relevance, conceptual rigor, and evidence transparency. Table III evaluates the entire corpus presented in Table II.

While most sources demonstrate high conceptual rigor, particularly standards and governance frameworks, quality assessment alone does not determine inclusion in the analytical synthesis.

A second-stage analytical selection was applied to identify sources relevant to governance-level conceptualization. Specifically, only sources that explicitly employ the term "metadata governance" or reflect governance constructs such as decision rights, accountability, and control mechanisms were included in the core analysis.

TABLE III. QUALITY ASSESSMENT AND ANALYTICAL CLASSIFICATION OF SOURCES

No	Source	Type	Gov. Relevance	Conceptual Rigor	Evidence Transparency	Score	Analytical Role
1	DAMA-DMBOK2 Revised (2024)	Industry Framework	High	High	High	3	Core
2	ISO/IEC 38500	International Standard	High	High	High	3	Supporting
3	COBIT 2019	Industry Framework	High	High	High	3	Supporting
4	ISO/IEC 11179	International Standard	Low	High	High	2	Contextual
5	ISO/IEC 19763	International Standard	Low	High	High	2	Contextual
6	ISO 19115	International Standard	Low	Medium	High	2	Contextual
7	ISO/IEC 27001	International Standard	High	High	High	3	Supporting
8	GDPR (EU 2016)	Regulation	High	Medium	High	3	Supporting
9	Perpres 39/2019	Regulation	High	Medium	High	3	Supporting
10	PDP Law (Indonesia)	Regulation	High	Medium	High	3	Supporting
11	Metadata Governance in Media Industry	Industry Study	Medium	Medium	High	2	Core
12	Atlan	Industry Source	Medium	Low	Low	1	Supporting
13	data.world	Industry Source	Medium	Low	Low	1	Supporting
14	Decube	Industry Source	Medium	Low	Low	1	Supporting
15	Dataversity	Industry Source	Medium	Low	Low	1	Supporting
16	DCAM v2.2	Industry Framework	Medium	High	High	3	Supporting
17	MetaStore	Industry Framework	Low	High	High	2	Contextual
18	COBIT-based IT Governance Metamodel	Academic Model	High	High	High	3	Supporting

This distinction ensures that quality assessment supports source credibility, while analytical selection determines conceptual relevance.

D. Governance-Level Conceptual Analysis

Based on the analytical selection, a subset of sources was examined in detail (Table IV). While Table III evaluates source quality and analytical roles, Table IV presents a governance-level interpretation of the same corpus to ensure traceability across the analysis. Table IV follows the same ordering as Table III to ensure traceability across the analysis. Selection is based on analytical relevance rather than quality score alone. These sources include governance frameworks, metadata standards, regulatory sources, practitioner-oriented materials, and conceptual references.

To distinguish governance from operational management, the analysis was grounded in governance theory, particularly ISO/IEC 38500 and COBIT, which conceptualize governance through decision rights, accountability, and the Evaluate–Direct–Monitor (EDM) cycle.

The analysis shows that while several sources employ the term "metadata governance," none fully satisfy formal governance criteria. Governance is generally embedded within metadata management practices rather than conceptualized as an independent governance system.

These findings indicate that metadata governance remains conceptually underdeveloped as a formal governance domain. The persistent conflation of governance with operational management reveals a structural gap in both literature and practice, reinforcing the need for clearer conceptual structuring and further research in metadata governance.

E. Conceptual Positioning of Metadata Governance

To synthesize the evidence landscape, Fig. 1 positions the reviewed sources along two analytical dimensions: 1) the degree of structural formalization, ranging from conceptual guidance to structured frameworks and further toward formal metamodel representations; and 2) the explicitness of metadata governance focus, ranging from implicit governance treatment to explicit conceptualization.

Within this positioning, DAMA-DMBOK2 Revised is situated in the region of conceptual guidance with a dominant management orientation. Industry and vendor-oriented sources exhibit more concrete and technically oriented structures, but generally lack explicit governance abstraction. In contrast, governance-oriented models, such as COBIT-based metamodels [44], demonstrate higher structural formalization but remain non-specific to metadata governance.

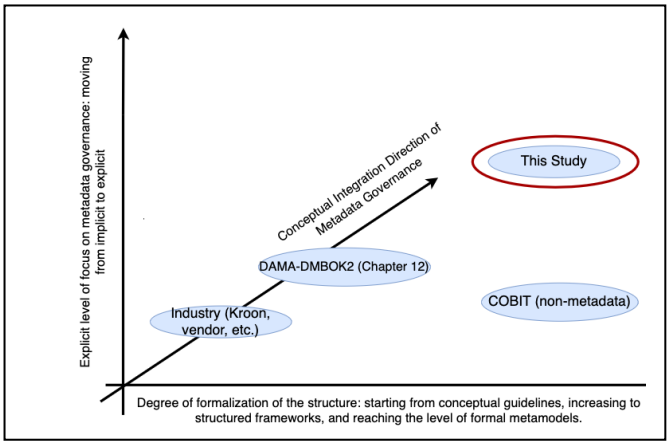


Fig. 1. Conceptual landscape of metadata governance research.

TABLE IV. GOVERNANCE-LEVEL CONCEPTUAL INTERPRETATION OF SOURCES

No	Source	Source Type	Explicit and Formal Governance?	Summary of Content	Governance Limitations
1	DAMA-DMBOK2 Revised 2024	Industry Framework	No (Embedded)	Mentions metadata governance within metadata management, covering standards, catalogs, lifecycle, and controls	No explicit EDM cycle; governance embedded in management
2	ISO/IEC 38500	International Standard	No	Defines governance principles (accountability, responsibility, EDM)	Not metadata-specific
3	COBIT 2019	Industry Framework	No	Defines governance domains and EDM cycle	Not metadata-specific
4	ISO/IEC 11179	International Standard	No	Defines metadata registry structures	No governance constructs
5	ISO/IEC 19763	International Standard	No	Defines metamodel for interoperability	No governance constructs
6	ISO 19115	International Standard	No	Provides domain-specific metadata structure	No governance constructs
7	ISO/IEC 27001	International Standard	No	Defines security governance controls	Not metadata-specific
8	GDPR (EU 2016)	Regulation	No	Defines compliance and accountability obligations	No governance structure model
9	Perpres 39/2019	Regulation	No	Defines national data governance obligations	No formal governance modeling
10	PDP Law (Indonesia)	Regulation	No	Defines personal data protection obligations	No governance modeling
11	Metadata Governance in Media Industry	Industry Study	No (Implicit)	Discusses governance in media workflows	No formal governance structure
12	Atlan	Industry Source	No (Practical)	Provides governance practices and tooling	Operational focus
13	data.world	Industry Source	No (Practical)	Focus on cataloging and governance tools	No formal governance abstraction
14	Decube	Industry Source	No (Practical)	Defines governance workflows	No institutional governance model
15	Dataversity	Industry Source	No (Educational)	Conceptual materials on metadata governance	No formal model
16	DCAM v2.2	Industry Framework	No	Defines data governance capability model	Not metadata-specific
17	MetaStore	Academic Framework	No	Defines metadata architecture	Technical focus only
18	COBIT-based IT Governance Metamodel	Academic Model	No (Non-metadata)	Formal governance metamodel	Not metadata-specific

Beyond these two axes, a third analytical dimension is introduced to represent the degree of conceptual integration. This dimension reflects how governance-related constructs are integrated into a coherent conceptual structure. These constructs include policies, roles, controls, accountability mechanisms, decision authority, lifecycle coordination, and assurance functions.

The analysis shows that most sources cluster in regions characterized by partial formalization and fragmented integration. Industry sources tend to provide localized operational integration without formal governance abstraction. DAMA-DMBOK provides partial integration by embedding metadata governance within broader data governance practices but does not establish it as an independent conceptual system. Formal governance metamodels exhibit strong abstraction but lack metadata specificity.

This study contributes by providing an integrated interpretation of governance-related constructs across heterogeneous sources, highlighting structural gaps and conceptual inconsistencies in metadata governance. Rather than proposing a design artifact, the study contributes by consolidating dispersed governance knowledge and clarifying the conceptual landscape of metadata governance.

V. Discussion

A. Answers to the Research Questions

This section synthesizes answers to the research questions based on the evidence presented in the Results.

1) RQ1 – Thematic coverage and research focal points: Research related to metadata governance is predominantly concentrated in technical and operational areas. Academic studies mainly address repositories, interoperability engineering, schema harmonization, metadata quality, and domain-specific applications, while explicit investigation of metadata governance as a standalone topic remains largely absent, as reflected by the zero eligible studies in the SLR. Metamodeling perspectives are rare and typically appear in interoperability standards rather than governance modeling. Grey literature broadens coverage by addressing stewardship, policies, security, and lifecycle concerns, yet these themes remain dispersed and lack consolidation into an integrated governance research stream.

2) RQ2 – Coverage of governance domains, processes, and conceptual components: Governance coverage across the literature is uneven and fragmented. Structural aspects of metadata are well documented in standards, while organizational constructs such as roles and policies are primarily addressed in frameworks and regulatory guidance. However, explicit articulation of governance domains, process structures, and conceptual relationships is largely absent. Most sources describe isolated elements without integrating them into a coherent governance architecture, reinforcing the lack of consolidated conceptual structures for metadata governance.

3) RQ3 – Evidence of validation practices: Reported validation practices remain limited. Academic studies occasionally employ technical validation (e.g., prototypes or performance evaluation), but validation of governance-oriented conceptual structures is rarely documented. Grey literature primarily relies on expert consensus and institutional review rather than systematic validation. Overall, the lack of explicit validation practices reflects the early and fragmented state of metadata governance as a research domain.

This limitation further reinforces the absence of formally defined and validated conceptual representations of metadata governance.

4) RQ4 – Conceptual and methodological deficiencies: The evidence reveals persistent deficiencies hindering the maturation of metadata governance as a research domain. These include limited integration between governance and formal modeling, inconsistent definition of governance constructs such as decision authority and accountability, and weak methodological rigor in validation. These deficiencies contribute to the absence of shared terminology and cumulative theoretical development.

While these findings provide a structured overview of the current state of metadata governance research, they also indicate a deeper conceptual limitation. The absence of governance-oriented abstraction across studies suggests that metadata governance is not yet recognized as a distinct analytical domain, but rather remains embedded within technical or operational perspectives.

B. Fragmentation of Metadata Governance as a Research Domain

The findings indicate that metadata governance has not yet evolved into a consolidated research domain. Academic studies largely approach metadata from technical perspectives, while governance aspects remain implicit or embedded within broader data governance discussions. Grey literature provides normative guidance through standards and frameworks, yet these sources remain heterogeneous and rarely integrated into coherent conceptual structures. This fragmentation reflects an imbalance between technical maturity and governance conceptualization in the metadata domain.

This fragmentation is not merely descriptive, but reflects a structural misalignment between the increasing strategic importance of metadata and the lack of corresponding governance formalization. As a result, the domain lacks conceptual coherence and cumulative theoretical development.

C. Governance Versus Management in Metadata Practices

A recurring pattern across the evidence is the conflation between governance and management. Many sources employ the term “governance” to describe operational controls, workflows, or tooling practices.

However, governance theory consistently distinguishes governance from execution. Institutional perspectives such as the G20/OECD governance principles frame governance as an arrangement of decision authority, accountability, and oversight structures. In the data domain, Khatri and Brown emphasize governance as the allocation of decision rights and responsibilities across critical decision areas, while DAMA positions governance as the function establishing accountability and policy direction for data assets.

This theoretical grounding is further operationalized in IT governance standards such as ISO/IEC 38500 and COBIT, which define governance through decision authority and oversight cycles (e.g., Evaluate–Direct–Monitor), clearly separating governance from operational management. However, analysis of the reviewed sources indicates that most references do not explicitly define governing bodies, decision domains, or governance cycles specific to metadata. Instead, governance is often implicitly reduced to management practices.

Clarifying this distinction is essential for advancing metadata governance as a formal governance domain grounded in decision authority and institutional accountability rather than operational best practices. This distinction is further supported by governance frameworks such as ISO/IEC 38500 and COBIT, which define governance in terms of decision authority, accountability, and oversight, rather than operational execution.

This inconsistency reveals a fundamental contradiction in the literature. While governance theory clearly distinguishes governance from management through decision authority and accountability, most metadata-related sources operationalize governance as execution-level practices. This contradiction contributes to conceptual ambiguity and limits the development of governance as an independent analytical construct.

D. Absence of Formal Conceptual Models and Metamodels

This observation is visually reflected in the conceptual positioning presented in Fig. 1. The review highlights a systematic absence of formal conceptual models and metamodels dedicated to metadata governance. While standards such as ISO/IEC 11179 and ISO/IEC 19763 provide structural rigor for metadata interoperability and registry architectures, they do not address governance abstraction. Conversely, governance metamodels developed in non-metadata domains demonstrate the feasibility of formalizing governance constructs but remain disconnected from metadata-specific contexts.

This disconnect explains the empty upper-right region in the state-of-the-art map, representing the intersection of explicit metadata governance focus, high structural formalization, and strong conceptual integration. Without formal conceptual representations, governance knowledge remains difficult to validate, reuse, or evolve cumulatively across studies.

This limitation has significant implications. Without formal conceptual models or metamodels, governance knowledge cannot be systematically validated, compared, or reused across different organizational contexts. This absence constrains both theoretical advancement and practical implementation.

It is important to clarify that this study does not propose a formal governance framework or metamodel. Instead, it synthesizes and consolidates existing knowledge to provide a structured conceptual understanding of metadata governance and to identify research gaps.

The objective of this study is not to design or validate a governance model, but to identify conceptual patterns, gaps, and inconsistencies in the current body of knowledge. This positioning aligns with the nature of multivocal literature review, which emphasizes synthesis rather than artifact development.

E. Implications for Research and Practice

From a research perspective, the lack of unified conceptual frameworks limits comparative analysis, validation, and theory development. For practice, organizations struggle to align technical metadata initiatives with institutional governance objectives, regulatory accountability, and cross-organizational coordination. Tool-centric approaches risk reinforcing operational silos without establishing robust governance structures. These findings indicate the need for structured conceptual foundations as a prerequisite for meaningful operational standardization and governance maturation.

These findings indicate that current approaches are insufficient to support governance maturity. Without a structured conceptual foundation, organizations risk implementing fragmented governance mechanisms that lack consistency, accountability, and strategic alignment.

F. Critical Synthesis and Limitations

The analysis reveals a clear problem–concept gap in metadata governance. While the need to govern metadata is increasingly evident in practice, the conceptualization of metadata governance remains underdeveloped and insufficiently formalized in the literature.

This gap is reinforced by three structural issues: 1) the dominance of technical and operational perspectives, 2) the fragmentation of governance constructs across heterogeneous sources, and 3) the absence of formal conceptual models supporting governance abstraction.

Furthermore, contradictions between standards-oriented and industry-oriented sources highlight a tension between formal governance design and practical implementation. Standards emphasize consistency and formalization, whereas industry practices prioritize flexibility and operational efficiency. The lack of integrative conceptual models prevents these perspectives from being systematically reconciled. This tension remains unresolved.

Nevertheless, the findings should be interpreted in light of the limitations of multivocal literature reviews. The inclusion of grey literature introduces variability in methodological transparency, and the absence of peer-reviewed studies limits comparative academic validation. Despite these limitations, the synthesis provides a structured conceptual foundation for future research.

VI. Conclusion

This study advances the understanding of metadata governance by consolidating the state of the art through a

multivocal literature review that integrates academic and authoritative grey sources.

The findings indicate that metadata governance remains conceptually fragmented, weakly formalized, and insufficiently validated as a scientific domain. Existing standards and professional frameworks do not yet provide a coherent governance architecture or a formal conceptual model.

The analysis identifies several critical gaps, including the absence of explicit governance abstractions, limited integration across governance domains, and weak validation practices.

These findings contribute to the development of a conceptual foundation for future theoretical, empirical, and methodological research on metadata governance.

A. Limitations of the Study

This multivocal review faces several limitations. First, the scarcity of peer-reviewed studies explicitly addressing metadata governance limits comparative academic evidence, reflecting the early maturity of the domain and motivating the inclusion of grey literature. Second, grey literature exhibits heterogeneous methodological transparency, which may affect replicability across review contexts. Third, the study emphasizes conceptual consolidation rather than empirical evaluation; therefore, findings primarily inform theoretical positioning rather than operational effectiveness.

B. Future Work

Despite the growing importance of metadata for interoperability and digital governance, the field remains conceptually fragmented and methodologically immature. The absence of consolidated theoretical foundations and formal models limits cumulative knowledge development and cross-context applicability.

Future research should elevate metadata governance beyond operational metadata management by framing it as an institutional governance problem involving decision authority, accountability, and sustainability. Future research may examine the development of governance frameworks, reference architectures, and conceptual metamodels for metadata governance constructs and relationships. Further research should also strengthen validation methodologies, explore governance–technology alignment, and conduct cross-domain and cross-sector comparative studies to understand how regulatory and institutional contexts shape metadata governance design.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by the Directorate of Research and Community Engagement (DRPM), Universitas Indonesia, through the PUTI Q2 2024–2025 research grant (Grant No. NKB-596/UN2.RST/HKP.05.00/2024).

VII. DECLARATION ON GENERATIVE AI

The authors acknowledge the use of generative AI to assist in improving the clarity, grammar, and structure of the manuscript. The content, analysis, and conclusions remain the sole responsibility of the authors.

REFERENCES

[1] M. Foulonneau and J. Riley, Metadata for Digital Resources: Implementation, Systems Design and Interoperability. Chandos Publishing, 2008. [Online]. Available: www.chandospublishing.com

[2] T. Grimstad and P. Myrseth, “Information governance and metadata strategies as a basis for cross-sector e-Services,” in eChallenges e-2010 Conference, IEEE, 2010. [Online]. Available: www.eChallenges.org

[3] J. Horodyski, Metadata Matters, First Edition. CRC Press, Taylor and Francis Group, 2022. doi: 0.1201/9781003189732.

[4] T. C. Redman, “Data quality must reads for researchers,” J. Decis. Syst., vol. 26, no. 3, pp. 203–206, 2017, doi: 10.1080/12460125.2017.1307621.

[5] J. Scott, Metadata: The Most Potent Weapon in This Cyberwar The New Cyber-Kinetic-Meta War Author. Institute for Critical Infrastructure Technology, 2017.

[6] ISO/IEC STANDARD 19115-1, “Geographic information — Metadata — Part 1: Fundamentals,” International Organization for Standardization, Geneva, 2014.

[7] ISO/IEC STANDARD 11179-1, “Information technology – Metadata registries (MDR) – Part 1: Framework,” 2023. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:11179:-1:ed-4:v1:en

[8] ISO/IEC STANDARD 11179-3, “Information technology-Metadata registries (MDR)-Part 3: Metamodel for registry common facilities,” Jan. 2023.

[9] ISO/IEC STANDARD 11179-5, “Information technology – Metadata registries – Part 5: Naming and identification principles,” Geneva, 2015.

[10] ISO/IEC STANDARD 19763-1, “Information technology-Metamodel framework for interoperability (MFI)-Part 1: Framework,” Switzerland, Jun. 2015. [Online]. Available: www.iso.org

[11] ISO/IEC STANDARD 11179-6, “Information technology-Metadata registries (MDR)-Part 6: Registration,” 2023.

[12] European Union, “General Data Protection Regulation (GDPR), Regulation (EU) 2016/679,” 2016. Accessed: Oct. 16, 2025. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

[13] ISO/IEC 27001, “ISO/IEC 27001:2022 – Information security, cybersecurity and privacy protection — Information security management systems — Requirements,” Geneva, 2022.

[14] Indonesia, Presidential Regulation No. 39/2019 on Satu Data Indonesia (One Data Indonesia). Jakarta, Indonesia, 2019. Accessed: Nov. 24, 2022. [Online]. Available: https://jdih.setkab.go.id/PUUdoc/175860/Perpres_Nomor_39_Tahun_2019.pdf

[15] Pemerintah Republik Indonesia, “Undang-Undang Nomor 27 Tahun 2022 tentang Perlindungan Data Pribadi,” 2022. [Online]. Available: https://peraturan.bpk.go.id/Details/229798/uu-no-27-tahun-2022

[16] R. W. Kroon and C. Crosby, “Media and entertainment metadata governance,” J. Digit. Media Manag., vol. 12, no. 2, pp. 143–156, Sep. 2023, [Online]. Available: https://www.eidr.org/assets/JDMM_12_2_JDMM0004_Kroon.pdf

[17] A. Prabhune, R. Stotzka, V. Sakharkar, J. Hesser, and M. Gertz, “MetaStore: an adaptive metadata management framework for heterogeneous metadata models,” Distrib. Parallel Databases, vol. 36, no. 1, pp. 153–194, Mar. 2018, doi: 10.1007/s10619-017-7210-4.

[18] OECD/ADB, “Digital government strategies and institutional frameworks in Government at a Glance Southeast Asia 2019,” Paris, 2019. doi: https://doi.org/10.1787/63029fcb-en.

[19] World Bank, “Toward a World-Class Labor Market Information System for Indonesia An Assessment of the System Managed by the Indonesian Ministry of Manpower,” 2021. Accessed: May 29, 2023. [Online]. Available: https://openknowledge.worldbank.org/server/api/core/bitstreams/906d1c88-4c68-5427-acd3-288aa4abd074/content

[20] World Wide Web Consortium, “Data on the Web Best Practices: W3C Recommendation,” W3C. Accessed: Oct. 15, 2025. [Online]. Available: https://www.w3.org/TR/dwbp/

[21] A. Yulfitri, D. Indra Sensuse, M. Bahrul Ulum, and Y. Fauzia Achmad, “Metadata Management to Accelerate Big Data Implementation,” Journal

of Informatics and Communications Technology, vol. 6, no. 2, pp. 109–117, 2024, doi: 10.52661.

[22] E. A. Purwanto, "Smart city as an upshot of bureaucratic reform in Indonesia," International Journal of Electronic Government Research, vol. 14, no. 3, pp. 32–43, 2018, doi: 10.4018/IJEGR.2018070103.

[23] OECD, "A data-driven public sector: Enabling the strategic use of data… OECD Publishing," May 2019. doi: 10.1787/09ab162c-en.

[24] Y. Wand and R. Weber, "Research commentary: information systems and conceptual modeling-a research agenda," Information Systems Research, vol. 13, no. 4, Dec. 2002, [Online]. Available: https://www.researchgate.net/publication/201169112

[25] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3," Engineering, vol. 45, no. 4ve, p. 1051, 2007, doi: 10.1145/1134285.1134500.

[26] Organisation for Economic Co-operation and Development, G20/OECD principles of corporate governance 2023. OECD Publishing, 2023. doi: 10.1787/fc7fe92b-id.

[27] OpenStax, Principles of management. OpenStax, Rice University, 2019.

[28] V. Khatri and C. V. Brown, "Designing data governance," Commun. ACM, vol. 53, no. 1, pp. 148–152, 2010, doi: 10.1145/1629175.1629210.

[29] ISO/IEC 38500, "ISO/IEC 38500:2024 – Information technology — Governance of IT for the organization," Geneva, 2024.

[30] ISACA, "COBIT® 2019 Framework: Governance and Management Objectives," 2019.

[31] ISACA, "COBIT® 2019 Framework: Introduction and Methodology," ISACA, 2019.

[32] DAMA International, DAMA-DMBOK2: Data Management Body of Knowledge (Revised Edition). Technics Publications, 2024.

[33] B. Otto, "A morphology of the organisation of data governance," 19th European Conference on Information Systems, ECIS 2011, no. June 2011, 2011.

[34] K. Weber, L. Cheong, B. Otto, and V. Chang, "Organising Accountabilities for Data Quality Management-A Data Governance Case Study.," Data Warehousing, no. October 2008, pp. 347–362, 2008, [Online]. Available: http://www.alexandria.unisg.ch/export/DL/Robert_Winter/204595.pdf

[35] J. Ladley, Data Governance How to Design, Deploy, and Sustain an Effective Data Governance Program, Second Edi. Academic Press Elsevier Inc., 2020.

[36] V. Garousi, M. Felderer, and M. V. Mäntylä, "The need for multivocal literature reviews in software engineering: Complementing systematic literature reviews with grey literature," in ACM International Conference Proceeding Series, Association for Computing Machinery, Jun. 2016. doi: 10.1145/2915970.2916008.

[37] V. Garousi, M. Felderer, and M. V. Mäntylä, "Guidelines for including grey literature and conducting multivocal literature reviews in software engineering," Inf. Softw. Technol., vol. 106, pp. 101–121, Feb. 2019, doi: 10.1016/j.infsof.2018.09.006.

[38] G. A. Bowen, "Document analysis as a qualitative research method," Qualitative Research Journal, vol. 9, no. 2, pp. 27–40, 2009, doi: 10.3316/QRJ0902027.

[39] EDM Council, "DCAM v2.2: Data Management Capability Assessment Model," 2022. [Online]. Available: https://edmcouncil.org/page/dcam

[40] Atlan, "Metadata Governance: Why You Shouldn't Neglect It." [Online]. Available: https://atlan.com/metadata-governance

[41] Data.world, "Metadata governance guide: Best practices, tools, and trends." [Online]. Available: https://data.world/blog/metadata-governance-guide/

[42] Decube, "Comprehensive Guide: Metadata Governance." [Online]. Available: https://www.decube.io/post/comprehensive-guide-metadata-governance

[43] Dataversity, "Metadata Governance: Crucial to Managing IoT." [Online]. Available: https://www.dataversity.net/articles/metadata-governance-crucial-to-managing-iot/

[44] J. Souza Neto and A. N. F. Neto, "Metamodel of the IT Governance Framework Cobit," Journal of Information Systems and Technology Management, vol. 10, no. 3, pp. 521–540, Dec. 2013, doi: 10.4301/s1807-17752013000300004.

Advanced Forensic Analysis Techniques for Insider Threat Detection in Database Management Systems

Kholod Saeed Talea AlQahtani, Mounir Frikha, M. M. Hafizur Rahman
Department of Computer Networks and Communications-College of Computer Sciences and Information Technology, King Faisal University, Al-Ahsa, Saudi Arabia

***Abstract*—This paper presents a real-time database forensic framework designed to detect insider threats within database management systems (DBMSs). Existing database forensic approaches, as identified through a systematic literature review employing the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology, are predominantly reactive and post-mortem in nature, lacking real-time SQL-layer visibility, forensic correlation, and evidence integrity assurance. To address these gaps, the proposed framework integrates native Microsoft SQL Server auditing mechanisms with a centralized ELK (Elasticsearch-Logstash-Kibana) pipeline to enable continuous evidence collection, automated correlation, and real-time visualization. The framework is evaluated against six insider-threat scenarios within the SPL ForensicDB—a simulated enterprise logistics database environment modeled after Saudi Post Logistics (SPL)—encompassing unauthorized access, data exfiltration, privilege escalation, data manipulation, backdoor creation, and audit suppression. Experimental results demonstrate high detection accuracy (precision: 0.92, recall: 0.88), a low false-positive rate of 3%, and alert latency consistently below five seconds, with minimal system overhead of 3.2% CPU utilization. The framework further ensures forensic integrity through SHA-256-verified, tamper-resistant audit logs and a structured chain-of-custody preservation mechanism compliant with ISO/IEC 27037:2012, making it suitable for both proactive security monitoring and legally defensible digital forensic investigations.**

Keywords*—Database management systems; insider threat detection; SQL server; forensic auditing; ELK stack; digital forensics

I. Introduction

The rapid development of digital technologies has fundamentally transformed how organizations store, manage, and secure data. A Database Management System (DBMS) is a software solution that enables users and applications to manage, retrieve, store, organize, and create data efficiently [1]. It serves as an interface between the database and its users, ensuring secure, consistent, and accurate information handling. DBMSs manage diverse user populations through concurrency control, data protection via security and integrity constraints, and backup-and-recovery mechanisms. Prominent examples include MySQL, Oracle, PostgreSQL, and Microsoft SQL Server [2].

The use of DBMSs has become central to contemporary organizations as the volume and sensitivity of stored information continue to grow. As data-driven operations have become the norm, the security and integrity of database environments have emerged as critical priorities. Insider threats—malicious or negligent actions by individuals with authorized access—are currently among the most harmful and difficult-to-trace categories of cyberattack [3][20]. According to the IBM Cost of a Data Breach Report [19], insider-related incidents account for a disproportionate share of total breach costs relative to external attacks, primarily because insiders operate within trusted boundaries and possess legitimate credentials.

This paper seeks to develop, deploy, and evaluate a real-time SQL-layer forensic auditing and insider-threat detection system built on Microsoft SQL Server, combining its native auditing capabilities with centralized log correlation and alerting. SQL Server is among a limited number of commercial DBMS platforms satisfying three core forensic requirements: tamper resistance, forensic auditability, and chain-of-custody preservation [4]. Research on database forensic readiness has shown that open-source DBMSs such as MySQL and PostgreSQL cannot support forensics without third-party extensions (e.g., pgaudit) [5], which cannot be considered tamper-proof or capable of preserving a defensible chain of custody essential for admissible forensic evidence. SQL Server, by contrast, provides SQL Audit and Extended Events, which produce binary, append-only, tamper-resistant .sqlaudit files satisfying the validation requirements of modern forensic audit research [6].

A. Research Gap and Motivation

Database-layer insider threats represent one of the most insidious forms of cyberattack because malicious insiders operate with legitimate credentials and SQL-level privileges that bypass conventional network- and OS-level security controls. Current database forensic solutions are predominantly post-mortem, disjointed, or based on outdated tracing mechanisms, rendering them ineffective at providing real-time visibility, evidence integrity assurance, or scalability to enterprise contexts. This paper addresses a specific and identified research gap: the absence of an authenticated, real-time, SQL-layer forensic pipeline capable of simultaneously supporting insider-threat detection, evidentiary integrity, and low operational burden within a production-scale database environment.

B. Novel Contribution

This paper proposes and empirically validates a unified SQL-layer forensic audit framework combining SQL Server Audit, Extended Events, and an ELK (Elasticsearch-Logstash-Kibana) pipeline to enable near-real-time insider-threat detection, correlation, and visualization. The contribution lies not in describing individual system components, but in demonstrating that native SQL Server mechanisms can be operationalized into a single forensic-ready pipeline capable of delivering:

- Real-time insider-threat detection at the SQL execution layer.
- Cryptographically verifiable evidence integrity via SHA-256 hashing.
- Performance overhead sustainable within enterprise production environments.

The remainder of this paper is organized as follows. Section II presents a systematic literature review of related work in database forensic auditing and insider-threat detection. Section III describes the experimental methodology, including the simulated environment configuration and attack scenario design. Section IV presents detection effectiveness and validation metrics. Section V addresses forensic integrity and evidence readiness. Section VI discusses operational and enterprise implications. Section VII summarizes the scientific contributions. Section VIII provides a comparative discussion of findings in relation to prior research. Section IX concludes the paper and outlines directions for future work.

II. Literature Review

A. Database Forensics and SQL-Layer Auditing

Research in database forensic auditing has evolved considerably, with scholars recognizing that classic OS-level artefacts are insufficient for investigating advanced insider attacks [6]. Early implementations relied on OS logs and general event logs, which lack the granularity necessary to reconstruct unauthorized DML operations, privilege escalations, or schema modifications at the SQL layer. Early research on log management [7] highlights that many trace formats lack inherent tamper resistance, rendering them vulnerable to manipulation by insiders and potentially inadmissible as forensic evidence. An alternative perspective based on external monitoring agents and third-party auditing platforms [8] introduces deployment complexity and may fail to capture engine-level execution traces.

Early work [9] identified that malicious activity at the DBMS layer frequently leaves implicit artefacts in transaction logs, cache buffers, and session metadata. However, these early contributions were largely confined to post-mortem analysis, permitting investigators to reconstruct incidents only after operational or financial damage had occurred. Khanuja and Adane [17] proposed a database forensic analysis framework that systematized the collection and analysis of SQL-layer artefacts, while Wagner [18] demonstrated DBMS forensic auditing using Extended Events (XEvents), establishing a technical precedent for the SQL-layer pipeline approach adopted in the present work. Wagner et al. [9] extended this research by examining database memory forensics and cache-pattern analysis for log verification, confirming that in-memory artefacts supplement binary audit logs in forensic reconstruction.

Studies on insider behaviour consistently report that such attacks unfold slowly and quietly, blending into routine operational tasks [10]. Behavioural profiling research [4][11] shows that insiders typically exhibit patterns such as accessing information beyond their assigned role, exporting sensitive data in bulk, or engaging in unusual after-hours activity. However, most proposed models analyse these behaviours only after the incident has occurred, relying on exported log files rather than real-time event streams [22][15]. This retrospective limitation creates a substantial detection gap, compelling investigators to reconstruct events after the fact rather than intervening during the active attack window.

Wei et al. [15] proposed an unsupervised anomaly detection scheme applied to database logs for proactive forensic investigation, demonstrating effectiveness on insider-threat scenarios; however, that approach lacked real-time pipeline integration and did not address cryptographic evidence integrity. Similar deep learning-based approaches for system log anomaly detection have also been explored in [24], further highlighting the potential of advanced analytics in forensic log analysis. Alzaabi and Mehmood [13] conducted a comprehensive review of machine learning methods for insider-threat detection, identifying precision and recall as primary evaluation metrics and underscoring the persistent challenge of high false-positive rates in anomaly-based systems. Manral and Somani [16] examined forensic capabilities in the presence of superuser insider threats, highlighting the critical importance of tamper-resistant logging mechanisms that cannot be disabled by privileged insiders—a requirement directly addressed by the SQL Audit binary format in the present study. Henriques et al. [14] surveyed forensics and compliance auditing for critical infrastructure, identifying ELK-based pipelines as among the most promising approaches for real-time log correlation and forensic visualization. Pachghare and Chopade [23] conducted a critical review of database forensics, identifying performance overhead as a key practical barrier to continuous auditing adoption, with enterprise acceptability generally requiring CPU overhead below 5%.

The theoretical foundation of this research rests on the Digital Forensic Readiness concept, which stipulates that systems should be designed to collect, preserve, and analyse evidence proactively rather than reactively [12]. Collectively, the reviewed literature—surveyed using the PRISMA methodology—supports the development of a unified SQL-based forensic pipeline combining real-time monitoring, automated correlation, and cryptographic evidence preservation.

B. Commercial Database Security Systems: A Comparative Overview

In addition to academic approaches, several commercial database security and auditing solutions have been widely adopted in enterprise environments. IBM Guardium and Imperva SecureSphere primarily rely on agent-based monitoring and network-level traffic inspection to detect suspicious database activities in real time [22]. While these systems provide strong detection capabilities, they often introduce substantial deployment complexity and may lack fine-grained forensic reconstruction at the SQL execution level due to their reliance on network-layer rather than engine-layer event capture.

Oracle Audit Vault focuses on centralized log collection and compliance auditing, offering a consolidated view of database activity across heterogeneous environments. However, its detection capabilities are generally less responsive in real-time attack scenarios compared to dedicated monitoring pipelines, and it does not natively provide cryptographic chain-of-custody mechanisms at the individual file level. In contrast, the proposed framework leverages native SQL Server auditing mechanisms combined with an ELK-based pipeline to provide near-real-time detection with minimal system overhead. The emphasis on tamper-resistant audit logs and SHA-256 cryptographic validation enables both operational monitoring and legally defensible evidence preservation, distinguishing the proposed approach as a lightweight, forensically robust, and practically deployable alternative within enterprise database environments [14].

III. Methodology

This research adopts a technical experimental approach to design and validate a SQL forensic auditing architecture using Microsoft SQL Server 2019 Enterprise. The experimental environment was constructed to simulate the operational characteristics of Saudi Post Logistics (SPL), incorporating realistic customer, financial, and audit datasets. A specialized forensic database—SPL ForensicDB—was designed to execute scripted insider attacks and capture the resulting evidence traces. The overall system architecture, as illustrated conceptually in Fig. 1, encompasses three phases: data collection at the SQL layer, processing and normalization via Logstash, and real-time alerting and visualization through Elasticsearch and Kibana.

The binary format of SQL Audit records enabled the use of sys.fn_get_audit_file() to transform audit streams into downstream-readable relational tables. This normalization step verifies the semantic consistency of logs prior to ELK pipeline ingestion. Filebeat was configured to continuously scan .sqlaudit and .xel directories and forward new events to Logstash, which performed parsing, enrichment, timestamp validation, and field normalization before indexing events in Elasticsearch. Kibana provided visualization dashboards enabling analysts to identify anomalies including bulk data extraction, repeated

authentication failures, suspicious privilege usage, and attempts to disable the auditing engine. The logical flow of the system is depicted in Fig. 2 and the overall architecture in Fig. 3.

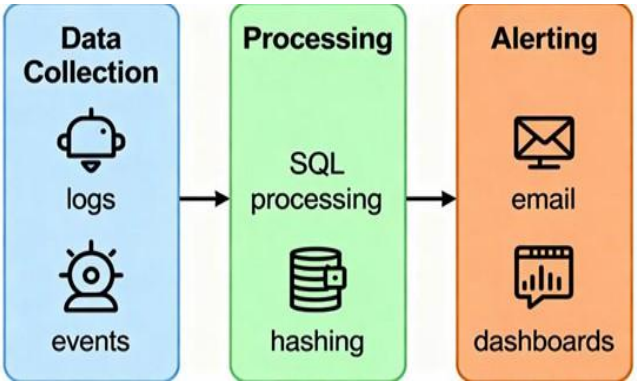


Fig. 1. Conceptual framework architecture showing data collection, processing, and alerting phases.

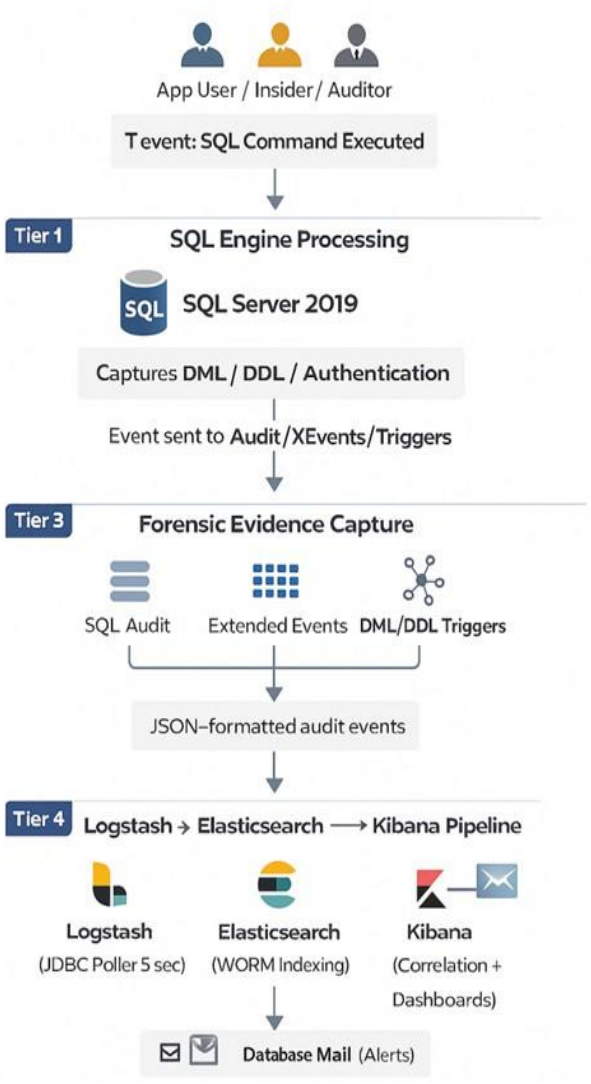


Fig. 2. Logical flow diagram.



Fig. 3. Forensic audit framework (system architecture).

Six insider-attack scenarios were scripted in T-SQL to evaluate the framework, as described in Table I. Performance was measured across five dimensions: precision, recall, false-positive rate, alert latency, and system performance overhead.

A. Experimental Environment Configuration and Log Sources

The framework was evaluated within a simulated enterprise operational environment modeled on Saudi Post Logistics (SPL) and running on Microsoft SQL Server 2019 Enterprise. The testbed incorporated realistic customer, financial, and audit data; RBAC structures; and controlled insider behaviour. Assessment employed a controlled experimental design in which insider attacks were executed as scripted T-SQL operations and evidence was ingested into the ELK pipeline from audit, trace, and trigger-based sources. The SQL Server instance was configured with secure authentication, structured log directories, and native audit mechanisms (Fig. 4). A dedicated SQL authentication login (SQLCONNECT) was created for forensic testing, with Mandatory Encryption enforced at the connection level.

System databases and the SPL ForensicDB are hosted within the instance (Fig. 5). SQL Server records system-level events and authentication information in the default ERRORLOG directory (Fig. 6). All forensic audit files generated by SQL Server Audit and Extended Events were stored in a dedicated directory at C:\ForensicAuditing\SQLAudit\ (Fig. 7). Two audit objects were

defined (Fig. 8): AuditServerToFile, capturing server-level events, and ForensicServerAudit, capturing database-level DDL, DML, failed logins, and privilege changes. These audits form the native SQL evidence stream prior to ELK ingestion.

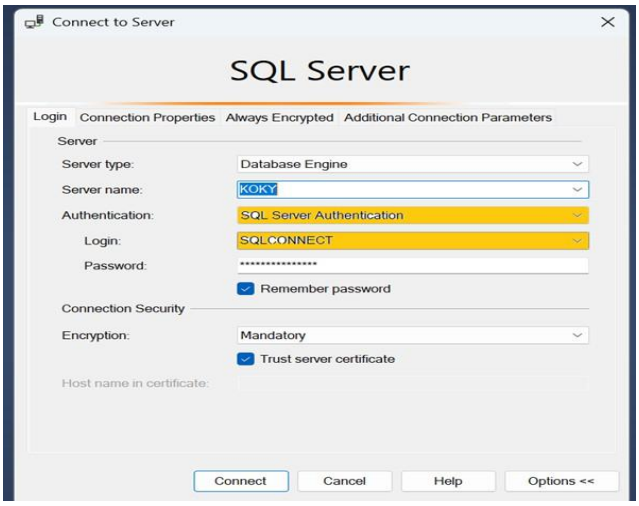


Fig. 4. SQL Server connection configuration showing SQL authentication and mandatory encryption.

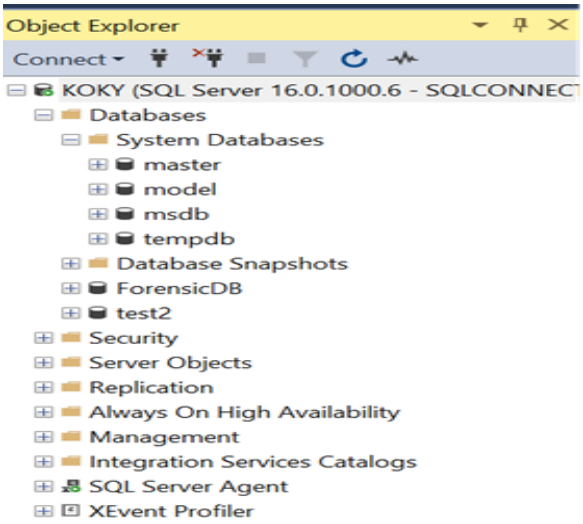


Fig. 5. SQL Server object explorer showing system databases and the forensicdb environment.

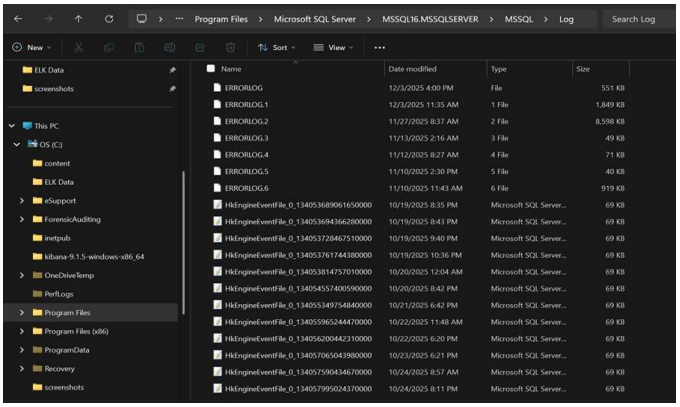


Fig. 6. Default SQL Server error log directory showing rotating ERRORLOG files.

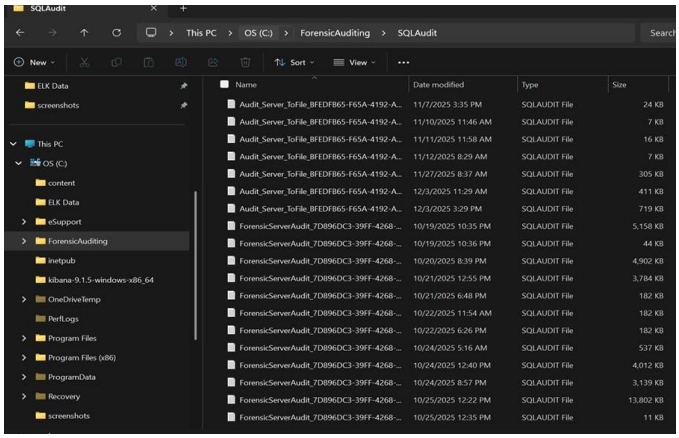


Fig. 7. SQL Audit directory containing generated .sqlaudit binary evidence files.



Fig. 8. Configured SQL Server Audits used for forensic event capture.

B. Simulated Database Environment: SPL ForensicDB

To ensure practical relevance, the SPL ForensicDB was constructed to mirror the schema complexity of a logistics and financial enterprise system. The Entity-Relationship Diagram (ERD) in Fig. 9 consists of four core modules:

- Human Resources Module: Employees (EmpID, Name, Department, ClearanceLevel); Users (UserID, Username, PasswordHash, RoleID).
- Customer Data Module: Customers (CustID, NationalID, Phone, Address, RiskScore)—containing PII data targeted in unauthorized-access and exfiltration scenarios.
- Financial Module: Transactions (TransID, Amount, Date, Sender, Receiver)—targeted for data manipulation scenarios.
- Audit Module: AuditLogMaster (LogID, EventType, TableName, OldValue, NewValue, ModifiedBy, Timestamp)—preserving before/after forensic attribution records via DDL/DML triggers.

The modules collectively replicate the real-world interdependencies between HR data, customer PII, and financial operations within SPL, ensuring that simulated attack scenarios are representative of genuine enterprise threat conditions.

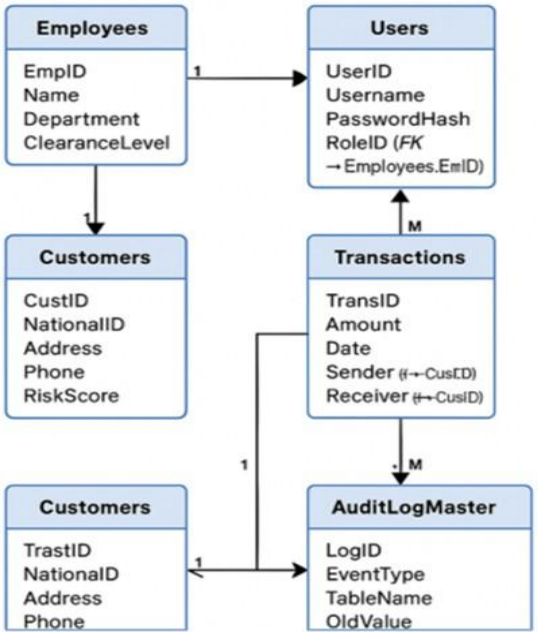


Fig. 9. ERD Relationship Diagram

C. Access Control and User Roles

To simulate realistic insider threats, the following user roles were defined:

- SysAdmin (SA): Full control, simulating a compromised administrator.
- DB Auditor: Read-only access to audit logs, representing the forensic investigator role.
- App User: Limited access via stored procedures, representing a normal employee.
- Malicious Insider: A user with db_datareader and db_datawriter privileges attempting unauthorized privilege escalation.

D. Insider-Threat Scenario Coverage

As summarized in Table I, six distinct insider-threat scenarios were designed to represent realistic and high-impact attack categories in enterprise database environments. Each scenario was engineered to test both detection accuracy and forensic completeness, ensuring that multilayered evidence could be reconstructed across SQL Audit logs, Extended Events traces, and DML/DDL triggers.

TABLE I. INSIDER THREAT SCENARIOS

Scenario ID	Insider-Threat Category	Description
S1	Unauthorized data access	Illicit reading of sensitive PII records
S2	Data exfiltration	Bulk extraction of customer data
S3	Privilege escalation	Unauthorized role and permission abuse
S4	Data manipulation	Financial record tampering
S5	Backdoor creation	Covert SQL login creation
S6	Audit suppression	Attempts to disable or tamper with audit logging

IV. DETECTION EFFECTIVENESS AND VALIDATION METRICS

Detection performance was quantitatively evaluated across all six scripted insider-threat scenarios. The framework successfully identified all six attack scenarios, with performance results summarized in Table II. These metrics confirm that the framework achieves high accuracy in detecting insider activity within the active attack window, enabling a timely forensic response without compromising production database performance. A post-execution review of detection logic, illustrated in Fig. 10, confirmed that all six attack sequences generated correlated alerts within the defined latency threshold, and no scenario evaded the pipeline.

TABLE II. SUMMARY OF DETECTION METRICS

Metric	Result	Forensic Meaning
Precision	0.92	Low false-alarm rate; investigators only review relevant alerts
Recall	0.88	High ability to detect real insider actions
False Positive Rate	3%	Minimizes alert fatigue and improves trust in system out-put
Median Alert Latency	2–3 sec	Enables near–real-time investigation of active insider at-tacks
CPU Overhead	3.2%	Meets <5% enterprise audit requirement without affecting production

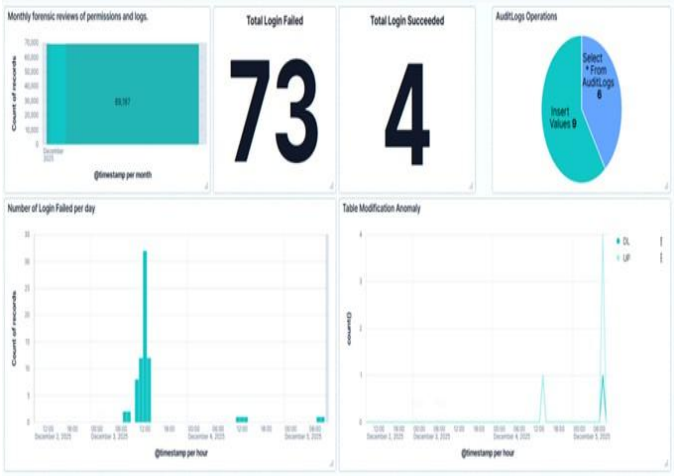


Fig. 10. Dashboard reflecting SQL audit events after malicious insider actions

V. FORENSIC INTEGRITY AND EVIDENCE READINESS

Beyond detection capability, the framework incorporates database-layer forensic readiness principles extending its utility to legally defensible evidence preservation. Upon acquisition, each evidence file was immediately processed to generate a SHA-256 cryptographic hash, consistent with the evidence integrity requirements of ISO/IEC 27037:2012 and the forensic integrity controls documented by Bush [6]. This process ensures that any post-acquisition tampering can be cryptographically detected, establishing an unbroken chain of custody from initial collection through analysis. The audit artefacts are stored as tamper-resistant binary .sqlaudit files with SHA-256 verification, NTP-synchronized timestamps, and WORM-like immutable indexing within Elasticsearch.

Table III presents the chain-of-custody validation results across all six evidence scenarios, confirming that all artefacts were successfully hashed, timestamp-synchronized, and alert-delivered within the defined latency bounds. Notably, Scenario S6 (Audit Suppression) validates the system's resilience against the most adversarial insider condition: an

attempt to disable auditing itself. The immutable Elasticsearch index confirmed in Table III for S6 demonstrates that even a forced audit engine shutdown generates a detectable and tamper-evident forensic record, addressing a critical gap identified by Manral and Somani [16]. This end-to-end evidence integrity makes the system applicable to legal and regulatory investigations—a requirement increasingly mandated under data governance frameworks [21].

TABLE III. CHAIN-OF-CUSTODY VALIDATION FOR INSIDER TEST EVIDENCE

Scenario ID	Evidence Captured	SHA-256 Ver-ified	Timestamp Synced	Output
S1	Customers read-perimeter breach	Yes	Yes	Email logged & delivered 3s
S2	Temp file exfiltration vector	Yes	Yes	Threshold breach alert 5s
S3	Unauthorized ALTER ROLE attempt	Yes	Yes	Role abuse visual flagged in Kibana
S4	UPDATE manipulation anomaly	Yes	Yes	0-value override detected & alerted
S5	Hidden login creation path	Yes	Yes	Backdoor login indexed & monitored
S6	Audit engine forced OFF	Yes	Yes	Immutable index lock confirmed in ELK

VI. OPERATIONAL AND ENTERPRISE IMPLICATIONS

The results confirm the operational feasibility of SQL-layer forensic auditing in enterprise settings. The co-existence of real-time monitoring with production workloads—demonstrated through a CPU overhead of 3.2%, well within the 5% enterprise acceptability threshold—substantially reduces the traditional trade-off between forensic comprehensiveness and operational performance. Centralized ELK-based visualization through Kibana (Fig. 10 and Fig. 11) enables analysts to rapidly identify anomalous patterns across temporal and behavioral dimensions, significantly reducing investigation time and analyst workload.

In practice, the framework provides organizations with an executable deployment roadmap for insider-threat detection, enhanced forensic preparedness, and improved accountability within database-centric infrastructures. The tamper-resistant binary audit format and SHA-256 chain-of-custody mechanism address a critical operational gap: the inadmissibility of forensic evidence from systems lacking verifiable integrity controls [7][16]. The framework thereby bridges the gap between security monitoring—a real-time operational function—and digital forensics—a post-incident legal function—within a single integrated pipeline.

VII. SUMMARY OF SCIENTIFIC CONTRIBUTION

The principal scientific contribution of this work lies in the empirical operationalization of a fully integrated SQL-layer forensic pipeline that bridges three previously disparate capabilities: real-time insider-threat detection, cryptographically verifiable evidence integrity, and sub-five-second alert latency within a production-representative database environment. Prior research has addressed these capabilities predominantly in isolation, limiting the practical applicability of individual contributions.

Database forensic research [17][18] demonstrated the evidential value of SQL-layer audit logs but did not operationalize them into real-time detection pipelines. Behavioural profiling studies [4][11] characterized insider-threat patterns effectively but lacked evidentiary integrity mechanisms required for legal admissibility. Evidence-integrity frameworks [7] established chain-of-custody methodologies but did not demonstrate integration with live SQL-layer detection at enterprise scale. The present work uniquely integrates all three capabilities within a single deployable pipeline and validates their simultaneous operation under realistic attack conditions.

The achievement of 3.2% CPU overhead during active forensic monitoring is particularly significant, empirically demonstrating that continuous SQL-layer auditing is operationally sustainable without dedicated hardware or substantial performance trade-offs—directly addressing the adoption barrier identified by Pachghare and Chopade [23]. The SHA-256 chain-of-custody mechanism, compliant with ISO/IEC 27037:2012, provides a legally defensible evidence trail that extends the system's applicability from operational security monitoring to formal digital forensic investigations. These combined properties represent a measurable advancement beyond descriptive system overviews and theoretical frameworks in the existing database forensics literature [5][12].

VIII. DISCUSSION

The findings confirm that the developed framework reliably captures insider activity at the SQL layer and provides forensically actionable insight. The framework successfully detected all six scripted insider-threat scenarios, capturing multilayered forensic evidence across SQL Audit logs, Extended Events traces, and DML/DDL triggers, with an overall precision of 0.92 and a recall of 0.88. The false-positive rate of 3% demonstrates that the detection logic effectively differentiates between genuine anomalies and normal operational activity. Alert latencies were consistently below five seconds across all scenarios, confirming near-real-time response capability sufficient for active-attack intervention.

The framework exhibited high forensic completeness. SQL Server Audit captured all applicable DDL, DML, and authentication events; Extended Events recorded engine-level data including bulk reads, query execution patterns, and performance anomalies; DML/DDL triggers preserved before-and-after values enabling precise forensic reconstruction of unauthorized changes. Kibana visualizations (Fig. 10 and Fig. 11) identified temporal spikes and abnormal access behaviours precisely corresponding to the simulated attack scenarios. Performance testing confirmed average CPU overhead of 3.2%, validating the feasibility of continuous forensic auditing in enterprise deployments and demonstrating the efficiency advantage of Extended Events over legacy SQL Trace mechanisms.

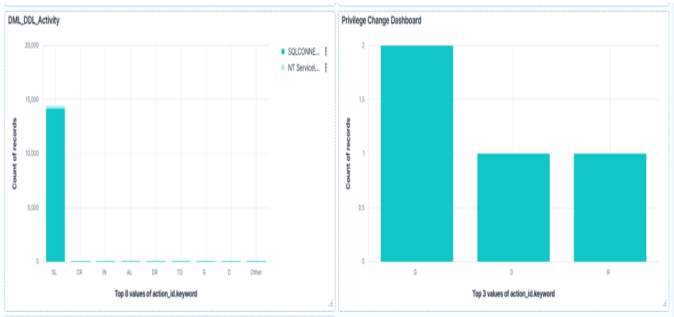


Fig. 11. Kibana privilege-change dashboard showing the impact of GRANT operations.

These findings are consistent with, and advance, prior contributions in the field. Wei et al. [15] demonstrated that unsupervised anomaly detection on database logs can support proactive forensic investigation; however, that approach lacked real-time pipeline integration and did not address cryptographic evidence integrity. The present framework advances this line of work by integrating detection with evidence preservation in a single operational pipeline. Manral and Somani [16] established the importance of forensic capabilities resilient to superuser manipulation—a property directly addressed through the tamper-resistant binary .sqlaudit format, which cannot be modified or deleted by privileged users without generating a detectable forensic artefact.

The false-positive rate of 3% compares favorably with behavioral profiling approaches reviewed by Alzaabi and Mehmood [13], which reported consistently higher false-alarm rates attributable to the difficulty of distinguishing malicious from legitimate privileged activity in purely behavioral models. The sub-five-second alert latency represents a substantial improvement over post-mortem forensic approaches [5][12], which offer no real-time intervention capability by definition. The 3.2% CPU overhead is consistent with—and below—the enterprise feasibility threshold identified by Pachghare and Chopade [23], who cited performance overhead as a primary barrier to continuous database auditing adoption. Henriques et al. [14] reached similar conclusions regarding the suitability of ELK-based pipelines for critical infrastructure forensics, further corroborating the architectural choices made in this work.

Compared with commercial solutions, the proposed framework offers complementary and, in several respects, superior properties for forensic use cases. Unlike IBM Guardium and Imperva SecureSphere, which rely on network-layer inspection and introduce substantial deployment overhead [22], the proposed system leverages native SQL Server engine mechanisms, minimizing infrastructure requirements while achieving greater SQL execution-level granularity. Unlike Oracle Audit Vault, which aggregates compliance logs rather than providing real-time detection, the ELK-based pipeline delivers sub-five-second alerting with a full forensic chain-of-custody. These distinctions position the framework as a practically deployable, forensically rigorous alternative suitable for resource-constrained enterprise environments seeking both security monitoring and legal evidentiary capabilities within a single system.

IX. Conclusion

This study has formulated, deployed, and thoroughly evaluated a SQL-layer forensic auditing framework that integrates SQL Server Audit, Extended Events, and an ELK-based evidence correlation pipeline. The analysis demonstrates that the proposed system achieves its design objectives across all evaluated dimensions: real-time insider-threat detection, cryptographic evidence integrity, and operational performance sustainability. The framework successfully detected all six scripted insider-threat scenarios with a precision of 0.92, a recall of 0.88, a false-positive rate of 3%, and alert latency consistently below five seconds, while maintaining CPU overhead of only 3.2%.

In addition to meeting functional and performance objectives, this work contributes to the theoretical understanding of how forensic readiness, behavioural analysis, and SQL-layer observability can be integrated into a coherent, production-deployable pipeline. It demonstrates that native database auditing mechanisms, when properly operationalized with modern analytics technologies, can deliver scalable insider-threat monitoring while satisfying the forensic integrity standards required for legal proceedings.

Future work will explore the extension of the framework to heterogeneous DBMS environments, the incorporation of machine learning-based anomaly scoring for adaptive threshold detection, and the addition of automated containment responses that not only detect insider attacks but also execute predefined countermeasures. Integration with Security Information and Event Management (SIEM) platforms and compliance with evolving data governance frameworks represent additional directions for broadening the framework's applicability.

Funding

This work was funded by King Faisal University, Al-Ahsa, Saudi Arabia KFU261953.

Acknowledgments

This work was supported through the Annual Funding track by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Al-Ahsa, Saudi Arabia, under Project No. KFU261953.

Conflicts of Interest

All authors declare no conflict of interest.

References

[1] K. D. Malakar, S. Roy, and M. Kumar, "Database management system: Foundations and practices," in Geospatial Technologies in Coastal Ecologies Monitoring and Management, Cham, Switzerland: Springer Nature, Jul. 2025, pp. 191–255.

[2] A. Tezuysal and I. Ahmed, Database Design and Modeling with PostgreSQL and MySQL: Build Efficient and Scalable Databases for Modern Applications Using Open Source Databases. Birmingham, U.K.: Packt Publishing, Jul. 2024.

[3] M. M. Alani, "Big data in cybersecurity: A survey of applications and future trends," Journal of Reliable Intelligent Environments, vol. 7, no. 2, pp. 85–114, 2021.

[4] P. Hakonen, "Insider threat detection using behavior analytics," in Detecting Insider Threats. Cham, Switzerland: Springer, 2022.

[5] A. Al-Dhaqm et al., "Digital forensics subdomains: The state of the art and future directions," IEEE Access, vol. 9, pp. 152476–152502, Oct. 2021.

[6] J. Bush, Practical Database Auditing for Microsoft SQL Server and Azure SQL. New York, NY, USA: Apress, 2022, ISBN: 978-1-4842-8633-3. DOI: 10.1007/978-1-4842-8634-0.

[7] E. N. Chukwuani and C. D. Ikemefuna, "Blockchain-based chain-of-custody models for tamper-proof evidence preservation in digital forensics investigations," 2023.

[8] A. M. Adebowale and O. B. Akinnagbe, "Cross-platform financial data unification to strengthen compliance, fraud detection and risk controls," World Journal of Advanced Research and Reviews, vol. 20, no. 3, pp. 2326–2343, 2023.

[9] J. Wagner, M. I. Nissan, and A. Rasin, "Database memory forensics: Identifying cache patterns for log verification," Forensic Science International: Digital Investigation, vol. 45, Art. no. 301567, Jul. 2023.

[10] N. Catrantzos, Managing the Insider Threat: No Dark Corners and the Rising Tide Menace. Boca Raton, FL, USA: CRC Press, Nov. 2022.

[11] O. K. Emmanuel, J. Aria, D. Jose, and C. Diego, Data-Driven Cybersecurity: How Analytics Can Predict and Prevent Insider Attacks, 2023.

[12] A. Selim and I. Ali, "The role of digital forensic analysis in modern investigations," Journal of Emerging Computer Technologies, vol. 4, no. 1, pp. 1–5, Dec. 2024.

[13] F. R. Alzaabi and A. Mehmood, "A review of recent advances, challenges, and opportunities in malicious insider threat detection using machine learning methods," IEEE Access, vol. 12, pp. 30907–30927, Feb. 2024.

[14] J. Henriques, F. Caldeira, T. Cruz, and P. Simoes, "A survey on forensics and compliance auditing for critical infrastructure protection," IEEE Access, vol. 12, pp. 2409–2444, Jan. 2024.

[15] Y. Wei, K. Chow, and S. Yiu, "Insider threat prediction based on an unsupervised anomaly detection scheme for proactive forensic

investigation," Forensic Science International: Digital Investigation, vol. 38, p. 301126, Oct. 2021.

[16] B. Manral and G. Somani, "Establishing forensics capabilities in the presence of superuser insider threats," Forensic Science International: Digital Investigation, vol. 38, p. 301263, Sept. 2021.

[17] H. K. Khanuja and D. S. Adane, "Database forensic analysis framework," Computer Science & Engineering International Journal, vol. 2, no. 3, pp. 27–41, 2012.

[18] J. Wagner, "DBMS forensic auditing via XEvents," in Proc. IEEE International Conference on Data Engineering (ICDE), 2018.

[19] IBM Corporation, Cost of a Data Breach Report 2024. IBM Security, 2024.

[20] N. F. Nassir, U. F. Rauf, Z. Zainol, and K. A. Ghani, "Revealing the multi-perspective factors behind insider threats in cybersecurity," Journal of Media and Information Warfare, vol. 17, pp. 65–82, Oct. 2024.

[21] B. Basic, P. Udovicic, and O. Orel, "In-database auditing subsystem for security enhancement," in Proc. 44th International Convention on Information, Communication and Electronic Technology (MIPRO), IEEE, Sept. 2021, pp. 1642–1647.

[22] G. B. Sadowski, K. K. Bimo, and T. Booth, "Critical capabilities for SIEM," Gartner Research Note, 2020.

[23] R. Pachghare and H. K. Chopade, "Critical review on database forensics," Digital Investigation, vol. 29, pp. 180–197, 2019.

[24] Z. Chen, J. Liu, W. Gu, Y. Su, and M. R. Lyu, "Experience report: Deep learning-based system log analysis for anomaly detection," arXiv preprint arXiv:2107.05908, Jul. 2021.

A Hybrid Explainable Ensemble Learning Framework for Health Risk Prediction

Hussain AlSalman
Department of Computer Science-College of Computer and Information Sciences,
King Saud University, Riyadh 11543, Saudi Arabia

***Abstract*—Early prediction of patients' health risk is a crucial component of the safe and effective implementation of clinical triage and timely intervention. Health risk data in the real world often tends to be small in size and limited, with class imbalance, making the overall accuracy and transparency of Machine Learning (ML) models insufficient and more difficult to achieve. This paper proposes a hybrid explainable ensemble learning framework for multi-class health risk prediction, which is built based on a stacking architecture with three strong base learners, namely Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). The XGBoost is chosen as the meta-learner due to its ability to learn complex non-linear probability mapping from the base models and provide complementary signals for the prediction. A complete preprocessing pipeline is implemented, covering missing data handling, systematic encoding of categorical variables, and strict separation between training and test sets to ensure unbiased assessment. Experimental results show that the proposed framework achieved an accuracy of 97.5%, which exceeds the accuracy results of individual models, which are 96.0%, 95.5%, and 96.0% for RF, XGBoost, and LightGBM, respectively. Additionally, the proposed framework integrates the predictive performance with the interpretable clinical decision support and transparency using SHapley Additive exPlanations (SHAP) method. The SHAP values are used to provide global and local explanations for revealing the most influential features that drive each prediction.**

***Keywords*—Health risk prediction; hybrid explainable; ensemble learning; meta-learner; LightGBM; SHAP**

I. Introduction

Accurate health risk prediction is an essential prerequisite for safe and efficient clinical decision-making [1]. In time-sensitive decision-making settings, clinicians faced the task of prioritizing patients' cases based on routinely available physiological measurements [2]. Conventional methodologies mostly rely on rule-based thresholds for integrating vital sign parameters and triggering escalation protocols [3]. A striking example is the National Early Warning Score 2 (NEWS2), which is widely implemented and standardized in clinical practice [3]. It explicitly uses oxygen supplementation and alternative calculation schemes to better incorporate respiratory risk into the normal clinical process of patient care [3]. Despite the fact that the Early Warning Score (EWS) is defined by simplicity, transparency, and ease of operationalization, its predictive efficacy differs from patient to patient and clinical environment to clinical environment [4]. Large-scale empirical studies have shown that NEWS2 is not better than its predecessor [4]. The NEWS, across the board, and, in some subgroups or in relation to some clinical endpoints, may in fact be shown to produce inferior predictive performance. Such constraints are the impetus for data-driven methodologies with the ability to unravel complex interrelationships of vital signs and clinical indicators [5].

In parallel, healthcare has seen a rapid growth in the number of machine learning (ML) methods with the goal of triage and deterioration prediction using routinely collected data during the medical examination and diagnosis phase [6]. Evidence from emergency departments' research shows that ML models are capable of exceeding conventional triage systems in predicting clinical outcomes [7]. This evidence presents the standard triage inputs, such as vital signs and basic clinical features, and provides support for the hypothesis that the process of algorithmic learners reduces mis-triage [7].

The reduction is achieved by identifying the interaction between features and non-linearity, which poses difficulties of representation within a deterministic scoring framework [7]. However, the effective use of models requires more than incremental increases in predictive accuracy. Clinical ML models need to be evaluated in realistic clinical operational conditions and should be designed with an adequate degree of interpretability. This process enables auditing, error analysis, and builds clinician trust. Hence, in addition to predictive accuracy, explainability is considered to be a very important criterion for therapeutic artificial intelligence [8]. Numerous high-performing models act as "black boxes" and interpretability layer takes on a particular importance in case of risk prediction [9].

A. Research Goals and Motivation

This study aims to develop an accurate and interpretable framework for multi-class health-risk level prediction from routinely available clinical indicators. The motivation for this work is two-fold. First, score-based systems, like NEWS2, are limited in the extent to which they use hand-crafted thresholds, additive rules, and are interpretable. This limitation may give rise to inconsistent discrimination across settings and endpoints [4].

Second, although traditional ML and single tree-based models provide powerful baseline models for structured prediction tasks [10], its application is sensitive to the sub-groups characteristics and may not be able to leverage the complementary strengths among learners fully. To overcome these shortcomings, an explainable hybrid ensemble framework is proposed to combine multiple base learners namely Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Light

Gradient Boosting Machine (LightGBM). The framework is able to learn an optimized combination policy by training a meta-learning, as is consistent with the theoretical underpinnings of stacked generalization. An important practical goal is to guarantee that performance improvements apply in realistic scenarios with limited and small sized imbalanced datasets. Accordingly, the developed framework focuses not only on overall accuracy but also macro-averaged performance measures.

Finally, to promote clinical transparency and decision-making, the framework is equipped with the SHapley Additive exPlanations (SHAP) explanation method. It can outline the features with the greatest impact at each risk assessment level using computed SHAP values [11]. Collectively, these goals align with a growing consensus that clinical machine learning applications should be both highly performant and interpretable to be credible and effective. In clinical triage and monitoring workflows, rapid risk stratification supports prioritization, escalation, and resource allocation decisions. Accordingly, the proposed model is designed as a decision-support tool that provides a risk-level estimate (Normal/Low/Medium/High) to assist clinicians rather than replacing clinical judgment.

B. Research Contributions

In the existing knowledge of data-driven clinical triage and vital-sign–based health risk classification, this work offers four main contributions, summarized as follows:

- An end-to-end, stacking-based framework is proposed that stacks multiple robust tree-based learners into a unified predictor for generalized, robust multi-class health risk level classification, addressing the problem of limited, small-sized datasets.
- Rather than a simple linear combiner, XGBoost is used as the meta-learner to learn a non-linear combination of base model probability outputs, thereby enabling recognition of cohort-specific integration patterns.
- Both general accuracy metrics and imbalance-aware metrics, such as macro-F1 scores and confusion matrices, are reported to ensure that improvements extend beyond the majority classes and remain meaningful across all risk classes.
- Global and instance-level explainability via SHAP is incorporated into the proposed framework for clinical transparency, audit, and feature-level reasoning.

The remaining part of the paper is organized into five Sections: Section II presents the related work on the considered problem and the analysis methods. Section III explains the research methodology, including the problem definition, the study dataset, building base learners, stacking the hybrid model, and model training and evaluation. Section IV presents the results and discusses the findings to justify the novelty and performance of the proposed framework's model. Section V illustrates the conclusions and future directions of current work.

II. Related Work

Health Risk classification using the rule-based Early Warning Score (EWS) remains in common use because of its easy formulation based on regularly documented vital signs measurements [3]. However, extensive research has shown that these scoring systems do not always improve discriminative performance across a wide range of outcomes and patient subgroups [4]. This limitation drives the movement toward more adaptable, data-driven approaches. Recent studies have continued to explore the use of ML methods for modeling and forecasting health risks, with a strong emphasis on the predictive ability of these systems [12]. These predictive systems have been used to enable intervention at early stages and preemptive decision-making, overcoming the limitations inherent in a conventional thresholding paradigm.

Kone et al. [13] introduced a study of health-risk prediction and prevention using ML methods integrated into protective healthcare pipelines. Multi-modal risk analysis has gained focus, with heterogeneous signals integrated to augment personalization and early alerting in a real-world context [14]. With the same goal and in the context of emergency department settings, triage models based on ML have been shown to perform better in terms of discrimination against clinically important outcomes [15]. Raita et al. [7] built ML models to predict Emergency Department (ED) outcome and compared them with classical triage methods and found that ML models can demonstrate better discrimination using routinely available prediction variables such as demographic data, vital signs, and presenting complaints. Likewise, Goto et al. [16] found a result of improved discrimination and reductions in both under- and over-triage among pediatric ED populations using ML-based prediction models compared to traditional triage tiers. Ensemble learning is one of the current approaches to increasing robustness in structured health data sets, especially when different constituent models exhibit complementary error dynamics [17].

Alternative representations to conventional tabular modeling have been analyzed to enhance usability for predicting health risks. Knowledge-graph-based methodologies have been used to aid risk classification by mining relationships among health concepts to facilitate an organized understanding of interconnected health entities [18]. Variants of deep learning architectures based on multi-layered perceptrons have also been explored for predicting health risk, hence the potential of representation learning in identifying complex patterns in medical data [19]. Workflow and infrastructure are huge factors in the real-world adoption of risk prediction systems. According to empirical investigations, adoption might be limited by insufficient integration with electronic health records (EHRs), which would result in deployment limitations that go beyond model correctness [20].

Explainability has become critical in the deployment of healthcare systems due to the need to explain complex ensemble predictions in risk-stratification situations [21]. Mukilan et al. [22] proposed a hybrid SVM-Naive Bayes ensemble with a combination of LIME and SHAP to achieve transparent air quality and health risk prediction. The authors showed that explanatory layers can be coupled with ensemble learners synergistically in order to increase interpretability [22]. Jayakarthik et al. [23] proposed an approach, called EnvHealthNet, a multimodal model for predicting environmental health risks. The approach follows the general

trend for integrating multiple sources of information. In maternal and antenatal contexts, ML-based risk-level prediction has also been investigated, and more recently, explicit modeling of explainability levels using techniques such as SHAP and LIME has been applied to explain maternal risk prediction. [24, 25]. Across studies and systematic reviews of ML in emergency department triage, the potential of clinical data and techniques to improve both consistency and accuracy in triage decisions is repeatedly highlighted, alongside existing issues such as a lack of datasets and a need for transparent evaluation and reporting [26, 27].

III. Methodology

This section presents the research methodology for developing the proposed framework for health risk prediction. Fig. 1 summarizes the complete workflow of the methodology. In the beginning, the study's dataset is described in detail. Then, the problem statement is formally defined to provide a foundation for the proposed solution.

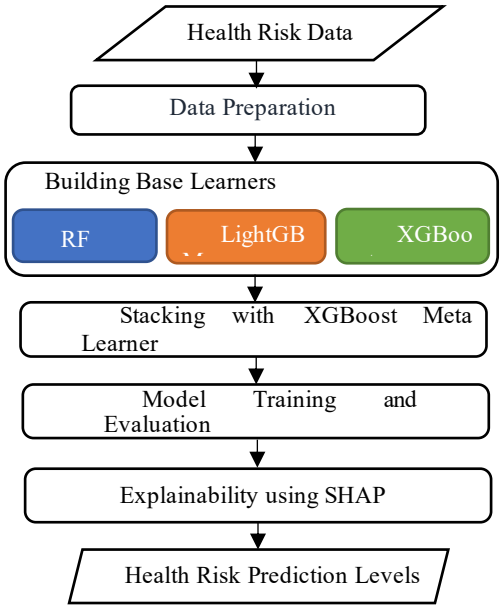


Fig. 1. Flowchart of research methodology.

Afterward, the preprocessing pipeline is described to handle missing values and categorical attributes consistently across both the training and inference phases. The base learners are then introduced, and the stacking ensemble is elaborated using XGBoost as the meta-learner to merge the probabilistic outputs from the base models. Next, the model training and evaluation process is applied to multi-class health risk levels. Finally, an interpretability layer based on SHAP is introduced to provide both global and instance-level explanations of the model decisions.

A. Problem Definition

The problem definition formulates the health risk prediction as a supervised multi-class classification task. Let $X \in R^{n \times d}$ denotes a feature matrix constructed from routinely recorded physiological indicators and clinical attributes. In addition, let $y \in \{1, \ldots, K\}$ represent the ground-truth risk class, in which $K = 4$ corresponds to {High, Low, Medium, Normal}. The goal is to learn a mapping $f: R^d \rightarrow \{1, \ldots, K\}$ that assigns the correct risk level to unseen cases. The risk classes represent ordered levels of severity relevant to real-world triage and monitoring, and the predictive performance should be effective with limited, small datasets, uniform across all classes, and not skewed towards the majority class.

B. Dataset of Study Design

The study uses an anonymized health-risk dataset consisting of variables based on vital signs and related clinical indicators, including metrics for oxygenation and a categorical variable for state of consciousness. The dataset consists of real-world patient data collected from Central Medical College and Hospital, with identifiers removed for privacy protection. It is publicly available on the KAGGLE platform under the name "Health Risk Prediction (Anonymized Real Data)" for conducting studies in health risk multi-class classification [28]. It contains 1,000 anonymized patient records with 9 features and a multi-class target with 4 risk labels: High, Low, Medium, and Normal. The dataset description and variable definitions are provided in Table I, while the class distribution is reported in Table II. We note that Patient_ID is used only for record indexing and is excluded from model training.

TABLE I. Description of Dataset Features

#	Column	Type (Role)	Description
1	Patient_ID	object (Identifier)	An anonymized patient record identifier (used for indexing; not a predictive clinical feature).
2	Respiratory_Rate	Integer (Feature)	Breathing rate (typically breaths per minute); reflects respiratory distress.
3	Oxygen_Saturation	Integer (Feature)	Peripheral oxygen saturation (SpO_2, typically %); indicates oxygenation status.
4	O2_Scale	Integer (Feature)	Oxygen scale/category used for scoring oxygenation (often aligns with clinical early-warning scoring scales).
5	Systolic_BP	Integer (Feature)	Systolic blood pressure (mmHg). It reflects the hemodynamic stability.
6	Heart_Rate	Integer (Feature)	Heart rate (beats per minute); indicates cardiovascular response/stress.
7	Temperature	float (Feature)	Body temperature (°C); indicates fever/hypothermia risk.
8	Consciousness	Object (Feature)	Level of consciousness (e.g., AVPU categories such as A/V/P/U); proxy for neurological status.
9	On_Oxygen	Integer (Feature)	Whether the patient is receiving supplemental oxygen (binary indicator, e.g., 0/1).
10	Risk_Level	Object (Target)	Multi-class health-risk label (Normal / Low / Medium / High) used as the prediction outcome.

TABLE II. DISTRIBUTION OF RISK LEVEL CLASSES

Class Label	Number of Instances
High	279
Low	255
Medium	306
Normal	160

C. Data Splitting and Preprocessing

A hold-out splitting strategy is used to randomly split the dataset into 80:20 training and test subsets to obtain a reliable estimate of generalization performance. The split is performed with a fixed random seed of 42 for reproducibility. The test subset is only for final evaluation and is not used to calibrate preprocessing or train meta-learners. Stratified partitioning is applied to ensure that imbalanced classes can be preserved across splits. This strategy is similar to real-world deployment scenarios in which models must handle unexpected data. Data preprocessing is then implemented as a standard pipeline to apply the same transformations during training and testing.

First, all features are checked for missing values. If numerical features contain missing values, median values are used to replace them. Also, missing values in categorical features are imputed using the mode. After that, categorical variables are converted to numerical values using one-hot encoding. It enables discrete clinical states to be represented without forcing a spurious ordinal relationship. The resulting transformation, z = T(x), is fit only on the training set and then used on both the training and test sets, ensuring that the evaluation is unbiased and that the deployed model will be able to handle new cases with the same preprocessing.

D. Building Base Learners

To capture heterogeneous patterns in structured clinical datasets, the proposed framework trains a diverse set of tree-based base learners at the first hierarchical level. Each base learner $m \in \{1, \dots, M\}$ induces a probability distribution over the K risk classes for a given transformed input $x \in \mathbb{R}^d$, given by $p_m(x) = [p_m(y = 1 \mid x), \dots, p_m(y = K \mid x)] \in \mathbb{R}^d$. In this study, powerful ensemble learners are built for tabular data type, especially bagging tree (RF), as well as gradient boosted decision trees (LightGBM and XGBoost). These ensemble learners can capture such nonlinear effects and feature interactions, which often appear in physiological measurements.

E. Stacking with XGBoost Meta-Learner

The proposed framework combines the base learners through stacked generalization implemented using a stacking classifier. In this configuration, the ensemble operates at the probability level, meaning that the inputs to the second-stage model are not raw features but the class-probability outputs produced by the first-stage learners. For an input instance x, each base learner returns a K-dimensional probability vector over the risk classes. Stacking mechanism constructs a meta-feature vector by concatenating these probability vectors across all base learners:

$$g(x) = concat(p_1(x), p_2(x), \dots, p_M(x)) \in \mathbb{R}^{MK} \quad (1)$$

Where M denotes the number of base learners, which is 3 in our framework. This representation captures complementary confidence patterns across models and provides a compact, information-rich input for second-stage fusion.

For the second stage, the framework uses XGBoost as the meta-learner, specified in the code as the final estimator. The meta-learner learns a nonlinear mapping from the stacked probability space g(x) to the final class probabilities:

$$p_{stack}(y \mid x) = h(g(x)), \hat{y} = arg \max_{k \in \{1,\dots,K\}} p_{stack}(y = k \mid x) \quad (2)$$

Using predicted probability of the stacked method, the meta-learner operates on a concatenation of base-model probability vectors $g(x) = concat(p1(x), \dots, pM(x))$. The XGBoost meta-learner learns a nonlinear fusion, $p_{stack}(y \mid x) = h(g(x))$, with $c\ v = 5$ used to generate unbiased meta-features during training.

To mitigate optimistic bias in meta-learning, the stacking construction specifies that the base-level probabilities used to train the meta-learner are generated via 5-fold cross-validation, which trains the framework to generate meta-features using cross-validated base predictions during fitting. We adopt stacking rather than fixed-weight aggregation because it learns a data-driven fusion rule from base-model probability outputs, enabling adaptive weighting across patient profiles. To validate this choice empirically, the proposed approach will be benchmarked against strong individual baselines using the same split, and both accuracy and macro-averaged metrics will be reported.

F. Model Training and Evaluation

The training phase is performed exclusively on the training subset. During fitting, the base learners are first trained on the training data, and the stacking mechanism internally constructs the meta-learning dataset as specified by the 5-fold cross-validation. Precisely, the training set is split into five folds; for each fold, the base learners are trained on the remaining folds and used to generate out-of-fold class probabilities for the held-out fold. These out-of-fold probabilities become the meta-features for training the XGBoost meta-learner, which learns the final fusion rule from base-model probability signals. After the meta-learner is trained, the base learners are refit on the full training set, enabling the final stacked model to use all available training data when producing predictions for unseen cases.

Evaluation is conducted strictly on the held-out test partition, which is never used during model fitting or meta-feature construction. After training, predicted labels are obtained using the trained pipeline of the base learners to produce $\hat{y}$. Performance is quantified using accuracy to summarize overall correctness and macro-F1-score to provide an imbalance-aware assessment that weights all classes equally. Accuracy, Precision, Recall, and F1- score are computed as:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (3)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (4)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (5)$$

$$F1\text{-}score = \frac{2\times(Precision\times Recall)}{(Precision+Recall)} \quad (6)$$

where, $TP, TN, FP,$ and FN are the number of true positives, true negatives, false positives, and false negatives of all classes.

No resampling like SMOTE, manual threshold shifting, or cost-sensitive reweighting is applied. To ensure fair assessment under imbalance, macro-precision, macro-recall, macro-F1-score, and class-wise results are reported in addition to accuracy. In other words, to confirm that performance is not dominated by the majority class, the macro-F1-score is computed by first calculating precision and recall for each class k, then averaging the per-class F1-scores as:

$$Macro\ F1\text{-}score = \frac{1}{K}\sum_{k=1}^{K} F1\text{-}score_k \quad (7)$$

In addition to scalar metrics, the confusion matrix is computed to inspect class-wise error patterns and identify systematic confusions between adjacent severity levels (e.g., Medium vs. High). The confusion matrix $C \in N^{K\times K}$ is defined as:

$$C_{i,j} = |\{x: y = i \wedge \hat{y} = j\}| \quad (8)$$

The confusion matrix counts how many instances of true class i are predicted as class j. Together, Accuracy, Macro-F1-score, and the confusion matrix provide a comprehensive evaluation: Accuracy captures overall correctness, Macro-F1-score emphasizes balanced performance under class imbalance, and the confusion matrix supports detailed clinical error analysis. In addition, the receiver operating characteristic (ROC) curve will be visualized to show the relationship between the true positive rate and the false positive rate at different decision thresholds.

G. Explainability Using SHAP

The methodology integrates SHAP explanations into the final stacked model, thereby enabling feature-level reasoning and clinical transparency. By decomposing a prediction into a baseline and contributions from various input variables, SHAP provides additive feature attributions. It is used to produce both local explanations and global explanations. Local explanations support the individual risk assignments by indicating the features that increase or decrease the probability of each class. For global explanations, mean absolute SHAP values specify the most important predictors in the ensemble learners group. The SHAP methodology gives coherent and computationally tractable explanations of tree-based models. By making the links between predictions and clinically significant signs, this interpretability layer aids in building trust, detecting unexpected model behavior, and simplifying the clinical review process.

IV. Experimental Results

In this section, the experiments of the proposed framework are conducted and discussed to provide in-depth interpretations of quantitative results as well as explanations for explainability analyses. The evaluation and findings are reported through three experiments. The first experiment presents the confusion matrix, ROC curves, class-specific measurements, and analysis to validate the predictive performance of the developed framework on a held-out test set. The second experiment compares the framework to individual base models to quantify the value of stacking and confirm that improvements stay consistent despite class imbalance. Finally, the third experiment demonstrates the results of interpretability integrated into the proposed framework. The SHAP values of global feature attributions are mapped to the interdependencies between data-level features and provide, at the same time, instance-level explanations for a representative test case. In the first experiment, the proposed hybrid ensemble is evaluated on the test set to demonstrate its performance in multi-class health risk classification. Fig. 2 shows the confusion matrix, which takes a class-by-class breakdown of correct predictions and misclassifications of the four risk classes.

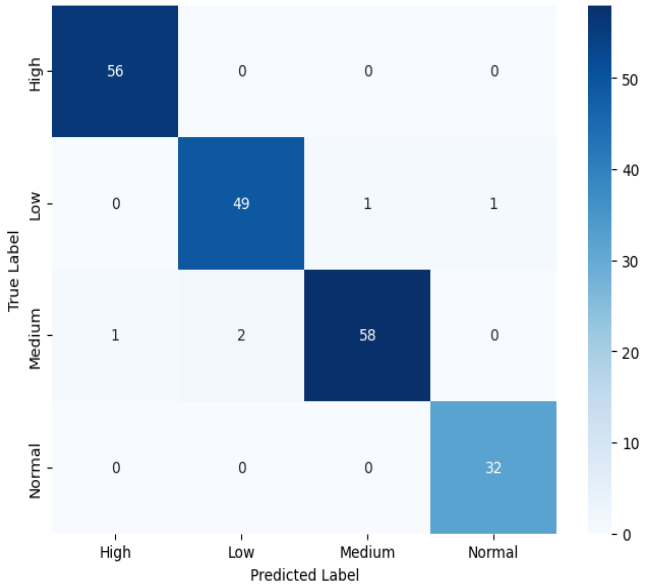


Fig. 2. Confusion matrix of the hybrid ensemble model.

As shown in Fig. 2, the main diagonal of the confusion matrix suggests that the model has a majority fit the observations to the correct risk categories. The corrected classified instances provide evidence of worthy overall discriminative performance. The model maintains high sensitivity for severe cases, which is an essential requirement for triage-oriented prediction. Following the confusion matrix, the results of other evaluation quantities are given in Table III, including the per-class precision, recall, F1-score, and the macro-averaged metrics.

TABLE III. Evaluation Results of the Hybrid Ensemble Model

Class Label	Precision	Recall	F1-score
High	0.9825	1.0000	0.9912
Low	0.9608	0.9608	0.9608
Medium	0.9831	0.9508	0.9667
Normal	0.9697	1.0000	0.9846
Accuracy	0.9750		
Macro avg.	0.9740	0.9779	0.9758
Weighted avg.	0.9751	0.9750	0.9749

As given in Table III, the obtained results support the confusion matrix findings. The model has an accuracy of 0.9750, along with a strong macro-average precision of 0.9740, macro-average recall of 0.9779, and macro-average F1-score of 0.9758. These results shaped the performance to be balanced in all classes. Furthermore, the high-risk class has a recall of 1.0000 and an F1-score of 0.9912, which implies that no severe cases are missed. The medium class shows a slightly lower recall of 0.9508 compared with the other classes, which is consistent with the low number of medium risk, classified as low or high risk, as shown in Fig. 2. However, the model's F1-score for the medium class is relatively high (0.9667), underlining good performance even for the most overlapped class. Fig. 3 shows the AUC-ROC plot of the hybrid ensemble model for each of the classes. The one-versus-rest ROC curve is obtained by treating each of the classes as the positive class, compared to the other classes at different thresholds of decision and the AUC is a threshold independent measure of discriminative capacity.

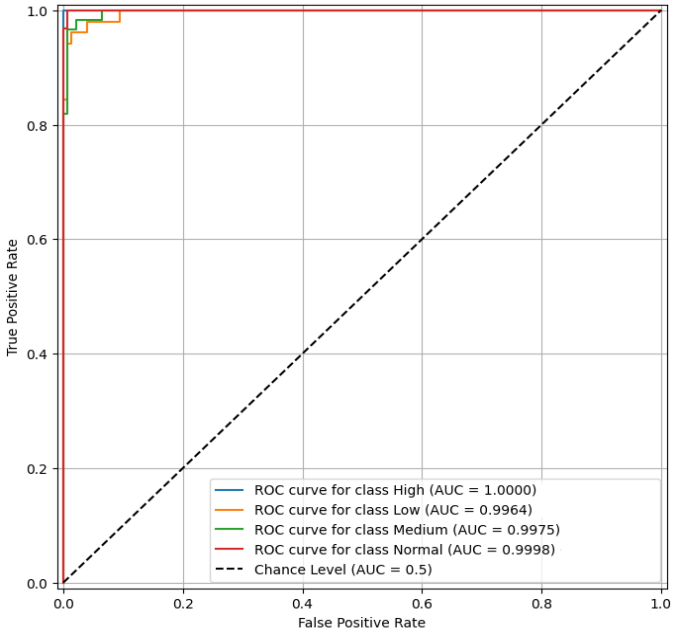


Fig. 3. One-versus-rest ROC curve of the hybrid ensemble model.

As shown in Fig. 3, the ROC curves for all classes in one-versus-rest classification are clustered in the upper-left quadrant, indicating that the classifier's discriminatory ability is respectable across all tested threshold values. The AUC scores are extremely high (1.0000 for High class, 0.9964 for Low class, 0.9975 for Medium class, and 0.9998 for Normal class). These results show that the ensemble model's probabilistic outputs achieve high separability between the risk categories and the remaining classes, supporting reliable threshold-based implementation when different operating points are required.

In the second experiment, the performance comparison between the hybrid ensemble model and single learners is documented in Fig. 4 and Table IV. Fig. 3 shows a comparison of the test set accuracy, and Table IV lists the macro-averaged precision, recall, and F1-score for each model. Fig. 4 shows that the hybrid ensemble model achieves the highest accuracy (0.9750), outperforming RF (0.9600), LightGBM (0.9600), and XGBoost (0.9550). This evidence identifies the framework as an approach for effectively utilizing the complementary strengths between the constituent base learners. The hybrid ensemble model improves generalization beyond that expected from any single robust model.

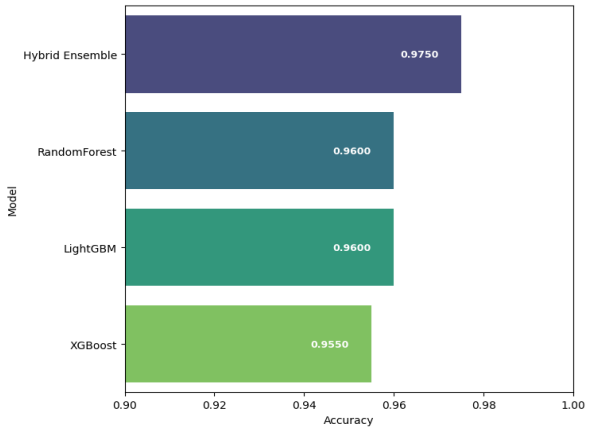


Fig. 4. Comparison of test accuracy between the hybrid ensemble model and individual baseline models.

TABLE IV. COMPARISON OF MACRO-AVERAGED EVALUATION METRICS BETWEEN THE HYBRID ENSEMBLE MODEL AND INDIVIDUAL BASELINE MODELS

Model	Macro F1-Score	Macro Precision	Macro Recall
RF	0.963	0.967	0.960
XGBoost	0.959	0.961	0.956
LightGBM	0.963	0.967	0.960
Hybrid Ensemble	0.976	0.974	0.978

To confirm that the majority class does not affect the improvement, Table IV presents the macro-averaged performance values. The hybrid mode achieved the best results, with a macro F1-score of 0.976, macro recall of 0.978, and macro precision of 0.974. In contrast, the individual models have lower scores. The RF and LightGBM have a macro F1-score of 0.963, whereas the XGBoost has a macro F1-score of 0.959. These results highlight that the hybrid model improved the balanced multi-class performance.

Furthermore, Table IV shows greater reliability for minority or critical classes than for overall accuracy alone. To validate the proposed model's results against the state-of-the-art method, a direct comparison with previously published methods on the same dataset is currently not possible, as no peer-reviewed benchmarks are available. Therefore, the results of the proposed model are compared against strong individual baselines on the same split and report macro-averaged metrics to account for class imbalance.

The third experiment evaluates the transparency of the framework using SHAP analysis. To provide global explanations of the model's predictions, SHAP summary plots are shown for each risk label, as shown in Fig. 5 to 8. These visualizations give each feature its relative importance in the aggregate model output. They show how feature values, whether high or low, influence the predicted probability towards or away from each class. Presentation of such summary plots by the risk

label is especially informative in multi-class settings since the same feature may contribute to divergent risk outcomes. This depends on the range in which its value falls, as well as its interactions with other variables. For example, a low oxygen saturation level may increase the probability of High-risk classification but also decrease the probability of Normal-risk classification.

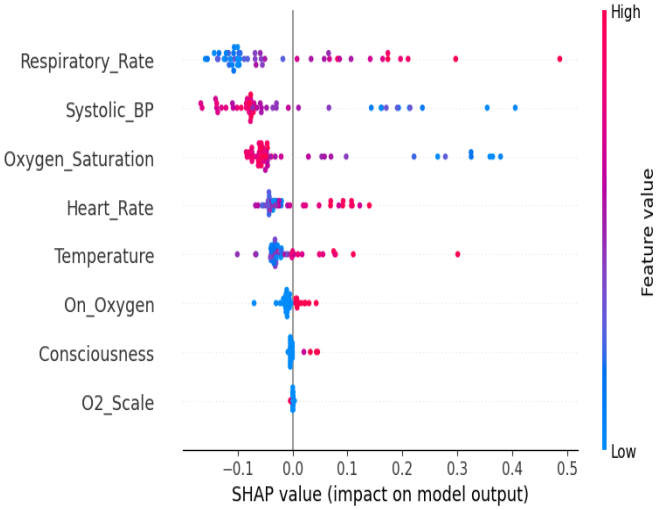


Fig. 5. SHAP summary plot for the high-risk class.

Fig. 5 gives the SHAP summary plot for the High-risk class. The most influential predictors, based on the selected variables, are respiratory rate and systolic blood pressure, followed in importance by oxygen saturation, heart rate, and temperature. This ordering agrees with known clinical expectations, in which respiratory decompensation and hemodynamic instability are the major indicators of high risk. The chromatic distribution further biases the model's output toward the High-risk category by applying pronounced values for features such as a high respiratory rate, low systolic blood pressure, and low oxygen saturation.

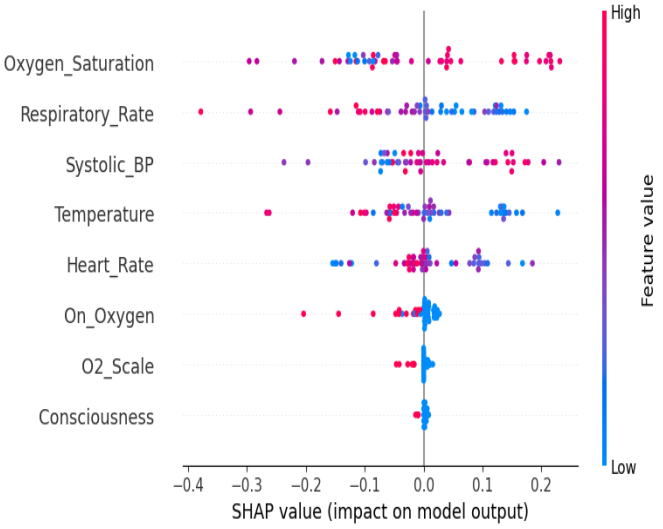


Fig. 6. SHAP summary plot for the low-risk class.

In Fig. 6, the SHAP summary plot for the Low-risk class shows the boundary of the features that influence the model's decision on Low risk, and how those feature value ranges modify the decision. The principal contributors mirror the clinically salient vital signs variables identified for each of the other classes, such as oxygen saturation, respiratory rate, systolic blood pressure and heart rate. However, their effects correspond with an intermediate spectrum of severity. In particular, the feature values reflecting mild deviations from physiology are associated with a higher probability of Low-risk. Those representing stable physiology will move the model's predictions toward Normal, and more obvious abnormal values will steer the model's predictions away from Low-risk toward Medium-risk or High-risk.

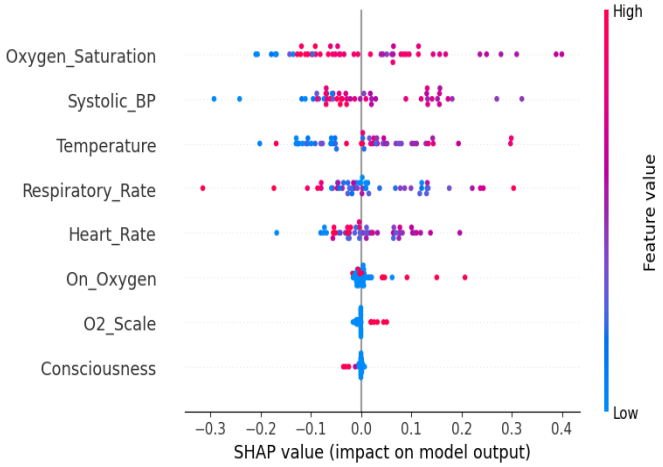


Fig. 7. SHAP summary plot for the Medium risk class.

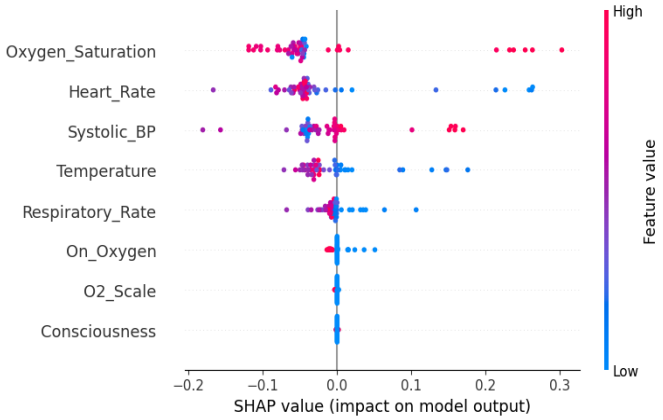


Fig. 8. SHAP summary plot for the Normal risk class.

Similarly, Fig. 7 and 8 represent the Medium-risk and Normal-risk classes, respectively. They show that oxygen saturation continues to have a dominant effect, whereas the contributions of heart rate, temperature, and respiratory rate vary across classes. The global explanations are consistent with the behavior of the confusion matrix. This consistency shows that misclassifications mostly concentrate near the Low and Medium risk border region, where physiological patterns can overlap. Together with Fig. 5, 7, and 8. Fig. 6 completes the class-wise global interpretability analysis and provides evidence that the model's Low-risk decisions are driven by clinically coherent

signals rather than spurious artifacts. For the dominant features, identified by SHAP, Fig. 9 visualizes their relationships using a pairwise scatter distribution plot. The points of the scatter distribution plot are colored by risk label to show class separation and overlap in feature space.

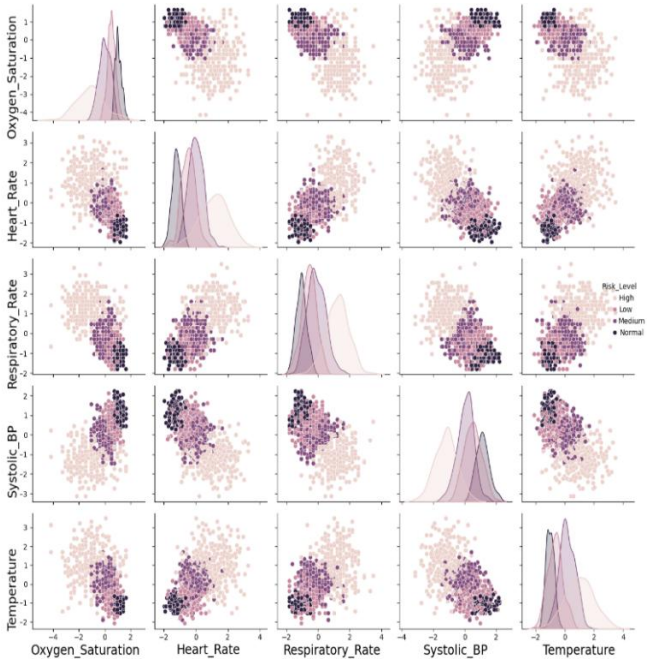


Fig. 9. Visualization of pairplot for the relationships between dominant features.

From Fig. 9, the scatter distribution of risk levels provides an empirical rationale for SHAP-based feature rankings. The visualization shows that the risk classes are in partially-separable regions of the combined feature space. For example, the separation between classes is increased once oxygen saturation is plotted side-by-side with respiratory rate and systolic blood pressure. This separation provides direct support for the prominence of these variables among the leading SHAP contributors. The overlap between classes, Low and Medium risk, provides a visual explanation for the constrained number of misclassifications described in Fig. 2, and the slightly reduced recall for the Medium class described in Table III.

In addition to global SHAP summaries, instance-level explanations are provided for a representative test case. The waterfall plots report feature contributions for each class (Normal, Low, Medium, and High), demonstrating not only why the model supports the Medium decision but also why competing classes are rejected. The local SHAP waterfall explanations for a specific test instance with true label Medium are given in Fig. 10 to 13, reporting class-wise contributions. Fig. 10 indicates that key features, notably systolic blood pressure, heart rate, temperature, and oxygen saturation, contribute negatively to the High class, preventing escalation to High-risk for this case. Fig. 11 highlights that respiratory rate strongly decreases the Low-risk probability, consistent with a physiological profile more concerning than the low class. Fig. 12 shows strong positive contributions from respiratory rate, systolic blood pressure, heart rate, temperature, and oxygen saturation, collectively increasing the probability toward the Medium class, which explains the final predicted outcome. Fig. 13 shows negative contributions from multiple vital-sign features, reducing the Normal class probability and justifying why the model rejects Normal for this instance.

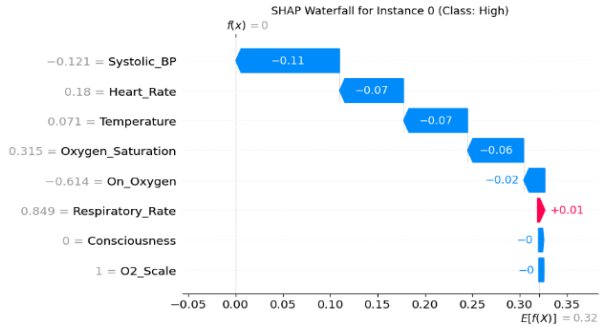


Fig. 10. SHAP waterfall plot for a specific test instance, with features influencing its probability of being classified as High risk.

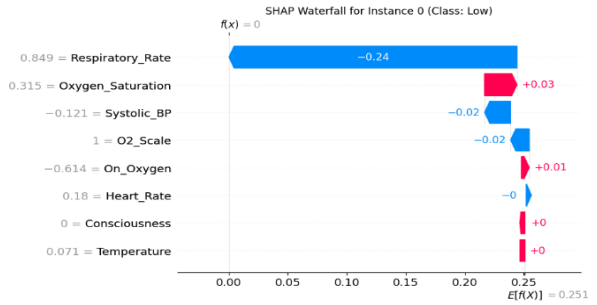


Fig. 11. SHAP waterfall plot for a specific test instance, with features influencing its probability of being classified as Low risk.

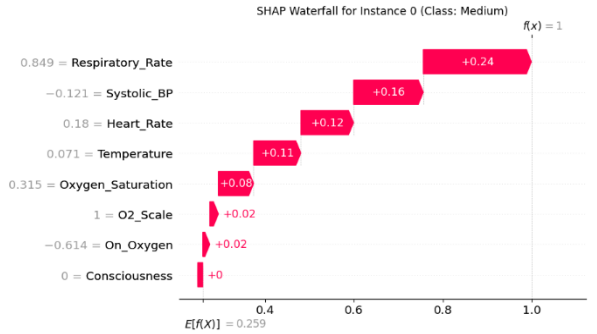


Fig. 12. SHAP waterfall plot for a specific test instance, with features influencing its probability of being classified as Medium risk.

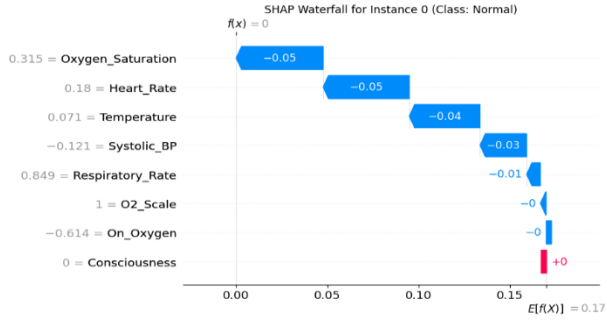


Fig. 13. SHAP waterfall plot for a specific test instance, with features influencing its probability of being classified as Normal risk.

Overall, these local explanations support the fact that the decision of the model is not only accurate but also traceable. As the same clinically relevant features identified at the global level in Fig. 5 to 8, Fig. 10 to 13 offer case-specific reasoning for providing cohesive transparency for clinical interpretation. These case-level attributions can support clinician review by identifying whether respiratory compromise (e.g., elevated respiratory rate, reduced oxygen saturation) or hemodynamic signals (e.g., lower systolic blood pressure) are the primary drivers of risk escalation. Across all experiments, the framework demonstrates the following advantages:

- Strong multi-class predictive performance with accuracy equal to 97.50%, and high per-class discrimination as demonstrated in Fig. 2, Fig. 3, and Table III.
- Consistent improvement over individual models in both accuracy and imbalance-aware macro metrics, as shown in Fig. 4 and Table IV.
- Clinically successful explanations through SHAP at both global and local levels, as visualized in Fig. 5 to 13.

These advantages support the proposed framework as an accurate and explainable framework for practical health-risk classification. For deployment, integration with EHR/HIS requires consistent data availability at the point of care, robust handling of missingness, latency-aware execution, and monitoring for dataset shifts over time. From an ethics and accountability perspective, the proposed model can be used as human-in-the-loop decision support, with auditability and explanation logs, subgroup performance checks, clinician override, and prospective validation before clinical adoption.

V. Conclusions and Future Work

This study proposed a hybrid explainable ensemble learning framework for multi-class health risk classification. The proposed framework is based on routinely collected clinical indicators, such as respiratory rate, oxygen saturation, blood pressure, heart rate, temperature, oxygen-related variables, and consciousness status. The hybrid classification model in the framework is built using the stacking ensemble concept to address a small, limited dataset and a class-imbalanced problem. The hybrid model generates synthesized complementary decision signals from multiple tree-based decision learners. It produces a singular, robust health risk classifier across four clinically significant classes: High, Low, Medium, and Normal.

Experimental results have shown that the framework demonstrates respectable predictive performance on a held-out test set, achieving 97.50% accuracy. More importantly, the model maintains high, balanced performance despite class imbalance, as shown by a high macro-level F1-score and valuable macro precision and recall values. The confusion matrix also shows that misclassifications are rare and mainly occur between adjacent severity levels, such as Low versus Medium. This outcome facilitates expected clinical overlap in borderline physiological presentations and is generally preferable to large severity jumps.

Beyond the predictive performance, the investigation highlighted the value of transparency and clinical interpretability. The SHAP-based explainability assessment provided both global and local insights into how the model operates. Global SHAP summary plots showed that variables of clinical importance, and particularly, variables reflecting respiratory compromise and physiological instability, such as respiratory rate, oxygen saturation, and systolic blood pressure, are predominant in risk discrimination. This interpretability provides evidence of the clinically plausible nature of the inferred decision function. The pairwise scatter distribution is a useful complement to these global explanations that revealed feature space configurations consistent with class separation patterns. The local SHAP waterfall visualizations provided patient-level reasoning for a representative, medium-risk case that captures all competing labels. Collectively, this body of evidence suggests that the framework has not only accuracy but also auditability. These interpretable justifications can be used to build clinician confidence and to provide an analysis of error in authentic decision-support contexts.

Because no peer-reviewed studies have reported machine-learning benchmarks on the dataset used in this study, this work provides the first academic baseline and a reproducible reference for future comparisons. A limitation of this work is that results are reported under a single hold-out split. Future work will include repeated runs with multiple random seeds and bootstrap confidence intervals to quantify performance uncertainty. Although this study adopts a stacking-based ensemble with an XGBoost meta-learner as the final configuration, other ensemble variants may be considered depending on the target operating requirements. For example, soft voting or weighted averaging can provide simpler aggregation. At the same time, alternative meta-learners, such as logistic regression or LightGBM, and different base-learner combinations may yield different trade-offs between performance, stability, and computational cost. A systematic comparison of these alternative configurations under the same data split and evaluation protocol is left for future work.

Moreover, future work will focus on validating the framework in external hospitals and care environments to quantify robustness to dataset shift and investigate subgroup performance across datasets. In addition, the probability calibration and the optimization of operating thresholds will be investigated to better align model outputs with the requirements of clinical decision-making. A second direction is to move the framework beyond its role as a snapshot prediction system towards use as more actionable decision support. This includes incorporating temporal trends from repeated vital sign measurements to enable earlier detection of deterioration and enrich features with other EHR signals. Finally, the SHAP layer will be improved with stability checks and clinician-facing explanation summaries, and then pilot or prospective evaluation to explore usability and impact within actual clinical workflows.

Acknowledgment

The author is grateful to the Deanship of Scientific Research at King Saud University's College of Computer and Information Sciences (CCIS) for funding this research.

References

[1] N. L. Edoh, V. M. Chigboh, S. J. C. Zouo, J. Olamijuwon, and Dentistry, "Improving healthcare decision-making with predictive analytics: A conceptual approach to patient risk assessment and care optimization,"

International Journal of Scholarly Research in Medicine, vol. 3, no. 2, pp. 1-10, 2024.

[2] L. Sandman and J. Liliemark, "Should severity assessments in healthcare priority setting be risk-and time-sensitive?," Health Care Analysis, vol. 31, no. 3, pp. 169-185, 2023.

[3] G. B. Smith, O. C. Redfern, M. A. Pimentel, S. Gerry, G. S. Collins, J. Malycha, D. Prytherch, P. E. Schmidt, and P. J. Watkinson, "The national early warning score 2 (NEWS2)," Clinical Medicine, vol. 19, no. 3, p. 260, 2019.

[4] M. A. Pimentel, O. C. Redfern, S. Gerry, G. S. Collins, J. Malycha, D. Prytherch, P. E. Schmidt, G. B. Smith, and P. J. Watkinson, "A comparison of the ability of the National Early Warning Score and the National Early Warning Score 2 to identify patients at risk of in-hospital mortality: A multi-centre database study," Resuscitation, vol. 134, pp. 147-156, 2019.

[5] N. S. Mosavi and M. F. Santos, "Enhancing clinical decision support for precision medicine: A data-driven approach," in Informatics, 2024, vol. 11, no. 3, p. 68: MDPI.

[6] R. Sánchez-Salmerón, J. L. Gómez-Urquiza, L. Albendín-García, M. Correa-Rodríguez, M. B. Martos-Cabrera, A. Velando-Soriano, and N. Suleiman-Martos, "Machine learning methods applied to triage in emergency services: A systematic review," International Emergency Nursing, vol. 60, p. 101109, 2022.

[7] Y. Raita, T. Goto, M. K. Faridi, D. F. Brown, C. A. Camargo Jr, and K. Hasegawa, "Emergency department triage prediction of clinical outcomes using machine learning models," Critical care, vol. 23, no. 1, p. 64, 2019.

[8] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and P. Q. Consortium, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," BMC medical informatics decision making, vol. 20, no. 1, p. 310, 2020.

[9] S. Niu, Q. Yin, J. Ma, Y. Song, Y. Xu, L. Bai, W. Pan, and X. Yang, "Enhancing healthcare decision support through explainable AI models for risk prediction," Decision Support Systems, vol. 181, p. 114228, 2024.

[10] M. Banerjee, E. Reynolds, H. B. Andersson, and B. K. Nallamothu, "Tree-based analysis: a practical approach to create clinical decision-making tools," Circulation: Cardiovascular Quality Outcomes, vol. 12, no. 5, p. e004879, 2019.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in neural information processing systems, vol. 30, 2017.

[12] F. Rahimian, G. Salimi-Khorshidi, A. H. Payberah, J. Tran, R. Ayala Solares, F. Raimondi, M. Nazarzadeh, D. Canoy, and K. Rahimi, "Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records," PLoS medicine, vol. 15, no. 11, p. e1002695, 2018.

[13] N. Kone, S. Singh, M. Noor, and N. Ranjan, "Health Risk Prediction and Prevention Using Machine Learning," in International Conference on AI Systems and Sustainable Technologies, 2025, pp. 101-113: Springer.

[14] A. B. Abraham, R. K. JS, N. V. Babu, and R. G. Nancy, "Multi-Modal AI Personal Health Risk Prediction and Early Alert System," in 2025 4th International Conference on Automation, Computing and Renewable Systems (ICACRS), 2025, pp. 775-782: IEEE.

[15] G. Feretzakis, A. Sakagianni, A. Anastasiou, I. Kapogianni, R. Tsoni, C. Koufopoulou, D. Karapiperis, V. Kaldis, D. Kalles, and V. S. Verykios, "Machine learning in medical triage: A predictive model for emergency department disposition," Applied Sciences, vol. 14, no. 15, p. 6623, 2024.

[16] T. Goto, C. A. Camargo Jr, M. K. Faridi, R. J. Freishtat, and K. Hasegawa, "Machine learning–based prediction of clinical outcomes for children during emergency department triage," JAMA network open, vol. 2, no. 1, p. e186937, 2019.

[17] L. R. Namamula and D. Chaytor, "Effective ensemble learning approach for large-scale medical data analytics," International Journal of System Assurance Engineering Management, vol. 15, no. 1, pp. 13-20, 2024.

[18] X. Tao, T. Pham, J. Zhang, J. Yong, W. P. Goh, W. Zhang, and Y. Cai, "Mining health knowledge graph for health risk prediction," World Wide Web, vol. 23, no. 4, pp. 2341-2362, 2020.

[19] T. Bikku, "Multi-layered deep learning perceptron approach for health risk prediction," Journal of Big Data, vol. 7, no. 1, p. 50, 2020.

[20] V. Sharma, I. Ali, S. van der Veer, G. Martin, J. Ainsworth, and T. Augustine, "Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records," BMJ health care informatics, vol. 28, no. 1, p. e100253, 2021.

[21] D. Saraswat, P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, P. N. Bokoro, and R. Sharma, "Explainable AI for healthcare 5.0: opportunities and challenges," IEEE Access, vol. 10, pp. 84486-84517, 2022.

[22] A. Mukilan, K. Abirami, K. Keerthishree, M. Srivani, and V. Vanitha, "Hybrid SVM-Naive Bayes Ensemble with LIME and SHAP for Transparent Air Quality and Health Risk Prediction," in 2025 13th International Conference on Intelligent Systems and Embedded Design (ISED), 2025, pp. 364-370: IEEE.

[23] R. Jayakarthik, D. Gopinath, M. A. Begum, A. D. Fuladi, A. Balaram, and C. Raja, "EnvHealthNet: A Multi-Modal Machine Learning Model for Commercial Environmental Health Risk Prediction," in 2025 5th International Conference on Pervasive Computing and Social Networking (ICPCSN), 2025, pp. 1121-1128: IEEE.

[24] F. Ahamad, "Detection of Antenatal Health Risk Level Using Machine Learning," in 2024 International Conference on Control, Computing, Communication and Materials (ICCCCM), 2024, pp. 426-429: IEEE.

[25] A. Rahman and M. G. R. Alam, "Explainable AI based maternal health risk prediction using machine learning and deep learning," in 2023 IEEE World AI IoT Congress (AIIoT), 2023, pp. 0013-0018: IEEE.

[26] J. S. Hinson, D. A. Martinez, S. Cabral, K. George, M. Whalen, B. Hansoti, and S. Levin, "Triage performance in emergency medicine: a systematic review," Annals of emergency medicine, vol. 74, no. 1, pp. 140-152, 2019.

[27] A. Naemi, T. Schmidt, M. Mansourvar, M. Naghavi-Behzad, A. Ebrahimi, and U. K. Wiil, "Machine learning techniques for mortality prediction in emergency departments: a systematic review," BMJ open, vol. 11, no. 11, p. e052663, 2021.

[28] R. Alam, "Health Risk Prediction (Anonymized Real Data)," URL https://www.kaggle.com/datasets/ludocielbeckett/health-risk-prediction-anonymized-real-data/data, [Last Access: 25-01-2026

A Digital Twin-Enabled Approach to Optimize Freight Fleet Operations in a Peruvian Transportation Company

Deyber Flores Cabezas, Liset S. Rodriguez-Baca
Professional School of System Engineering, Cesar Vallejo University, Lima, Peru

***Abstract*—The study evaluates the impact of a system enabled by Digital Twin on the optimization of the operations of a cargo fleet belonging to a transport company in Lima, Peru, during the year 2025. The proposed solution integrates OBD-II sensors, a Python processing engine, and a MongoDB-based data layer to build a synchronized virtual representation of 25 operating vehicles. An experimental, pre-experimental design with measurements before and after the intervention was applied to analyze changes in the frequency of accidents, the use of load capacity, and the monthly number of trips. The results show significant improvements: the frequency of accidents decreased from an average value of 0.08 to zero; the use of load capacity increased from 31 to 35 units; and the number of trips required to transport equivalent volumes decreased. These findings suggest that Digital Twin–based systems can support safer, more efficient, and data-driven operations in emerging logistics environments.**

***Keywords*—Digital Twin; fleet management; IoT; transportation; operational optimization**

I. Introduction

Freight forwarding companies are operating in rapidly changing conditions today. The pressure to comply with stricter schedules, maintain operational security, and make better use of resources requires constant coordination between people, vehicles, and processes. However, in many regions and more frequently in emerging economies, decisions continue to depend on partial systems or fundamental digital tools, which limit the ability to foresee contingencies or adjust operations in time [1], [2]. When the environment becomes more dynamic, and cargo volumes increase, these limitations translate into delays, greater wear and tear on the units, and a reduced margin to react to unforeseen events [3], [4].

In recent years, the idea of virtually representing a physical process through a Digital Twin (DT) has gained prominence as an alternative to face these challenges. A DT combines real-time data with models that enable interpretation of a system's state and anticipation of its behavior. Thanks to the integration, it is possible to observe the performance of a vehicle, a production line, or even an entire logistics network in a virtual environment that reflects what happens on the physical plane [5], [6], [7]. The technology, applied correctly, facilitates the early detection of failures and allows exploring how a system would respond to different scenarios before executing them in real operation.

While most studies have been conducted in highly automated facilities or in projects linked to autonomous driving, the principles underpinning DTs can also be applied in traditional fleets operating with modest technological resources [8], [9], [10].

In the transport of goods, having timely information on the vehicle's driving, mechanical condition, and use can make the difference between an efficient service and an expensive or risky one. When control is based solely on manual observations or incomplete reports, it becomes difficult to detect driving patterns that can lead to incidents, identify route deviations, or assess whether the load is being used properly. A DT offers a way to overcome these limitations by collecting continuous telemetry and converting it into indicators that support operational decision-making.

Despite this potential, there is still little evidence on DT implementations in small and medium-sized transport companies in Latin America. Many of them work with tight budgets and limited digital infrastructure, so it is essential to explore lightweight, cost-effective, and open component-based solutions that can be tailored to their needs. It is necessary to show that this type of technology can generate added value in safety and operational efficiency without requiring a high investment.

To address this need, a DT was implemented in a Peruvian freight transport company. The IT solution combines OBD-II sensors, Python microservices, a document-oriented database, and a real-time monitoring panel. Its effect is measured using an experimental, pre-experimental design, with three key indicators: accident frequency, load capacity use, and the number of trips required to meet demand. Taken together, the results point to meaningful improvements across all indicators and suggest that an accessible DT architecture can help strengthen safety and operational efficiency, even in companies with limited technological resources.

The article is organized according to the following sections: Section II summarizes the main contributions; Section III contains linked antecedents; Section IV explains the methodology and architecture of the IT solution; Section V presents the findings of the research; Section VI presents the discussion; Section VII presents the conclusion, and Section VIII outlines future work.

II. Research Contribution

The study identifies three aspects that help explain how Digital Twins (DT) can be applied in transport companies operating with limited technological resources.

- The first contribution is a DT architecture that operates without costly equipment or complex technological platforms. The proposal for the study uses affordable OBD-II sensors, open-source tools, and microservices that operate in a modular way, thereby avoiding reliance on specialized technological equipment.
- The study explores the effect of Digital Twin technology on the day-to-day performance of a freight transport company. While Digital Twins are already common in industrial and automation environments, their application in traditional fleets remains less studied. For that reason, three key indicators—accident occurrence, load capacity utilization, and number of trips—were compared before and after the implementation of the IT solution. The results made it possible to observe how continuous monitoring and data-driven analysis derived from the DT can contribute to safer and more efficient fleet operations.
- The study proposes a simple methodology. The methodology combines data capture via IoT devices, continuous information processing, and a set of indicators that can be adapted to different types of organizations. In addition to addressing the needs of the company analyzed, the framework serves as a guide for researchers and others interested in implementing DT in SMEs.

These contributions show that Digital Twins can be effectively applied in the transport sector, even in organizations that have had difficulty adopting more advanced technologies. The results obtained show that, even with low-cost solutions and reduced scale, it is possible to achieve clear improvements in operational performance. These results also open the door to new lines of research aimed at refining lightweight architectures and exploring their application in broader and more complex scenarios.

III. Related Work

In recent years, having a digital replica of a physical system has come to the forefront across various areas of engineering. The Digital Twin approach makes it possible to understand how a real process may behave and to anticipate possible operational situations. Although its use has been consolidated in industries with high levels of automation, in cargo transport, it is still in an exploratory stage, where companies seek to adapt their principles to the particularities of their operations.

Wang et al. [13] presented a vehicle navigation prototype based on a Digital Twin model, which improved the accuracy of automated decisions. The authors showed that by maintaining a constant flow of data between the vehicle and its virtual representation, it is possible to anticipate potentially dangerous maneuvers and correct them before they lead to an accident. That conclusion is relevant to the present study because it highlights an important point: when information is analyzed in real time, it becomes a valuable resource for reducing unwanted operational events.

Tang et al. [14] emphasize that Digital Twins must adapt to the environment in which they are implemented. According to their analysis, a generic architecture does not always yield reliable results, as each domain has specific conditions that must be reflected in the virtual model. Such an observation is especially relevant in cargo transport, where operating conditions change frequently and vehicles do not always have the same equipment or receive the same level of maintenance.

It is also relevant to mention the contributions from the industrial field. Ramasubramanian et al. [15] studied how Digital Twins support collaboration between operators and robots in a production plant. Although their work is not focused on logistics, they use Digital Twins to create a virtual representation that enables them to analyze different operational alternatives without interrupting the real process. Such an approach reduces both the risks and the costs associated with direct experimentation in the physical environment.

In the logistics field, Abdullahi et al. [16] developed a framework to improve sustainability by combining Digital Twins with real-time operational data. Their findings showed that increased digitalization enables more precise route planning, avoiding unnecessary travel and reducing both fuel consumption and emissions. A closely related concern in the study is the optimization of the number of trips required for a fleet to meet the same level of demand.

Madusanka et al. [17] reviewed the use of Digital Twins in the maritime domain and identified emerging trends related to monitoring, simulation, predictive analysis, and operational decision-making. Although their study focuses on maritime environments, its findings are relevant to freight fleet management because both contexts require continuous synchronization between physical assets and virtual models to improve safety, efficiency, and resource utilization.

Finally, Gerlach et al. [18] analyzed the use of "supply chain twins" to support the management of supply chains. Their study shows that by having a digital replica of each stage of the process, organizations can more clearly understand how the whole behaves and make informed decisions. This perspective supports the idea that a Digital Twin, even in smaller-scale operations or with limited resources, can deliver tangible improvements in efficiency and safety.

The revised precedents coincide in two central aspects. On the one hand, they argue that Digital Twins are a powerful tool for optimizing operational performance. On the other hand, they highlight the lack of research conducted in real-world settings, especially in environments with limited technological infrastructure. The present study addresses that gap by testing a simple, low-cost architecture in a Peruvian freight transport company.

IV. Methodology

A. Study Design

The study adopted a quantitative approach and an applied purpose, aiming to determine how a technological intervention could influence the operational performance of a freight

transport company. A pre-experimental design was used, with measurements taken before and after the system's implementation. A design of this kind is appropriate when the goal is to examine changes associated with an innovation without full control over the external variables that affect daily operations.

Under this design, it was possible to directly identify the effect of the Digital Twin on indicators used by the company in its usual management, without altering the real context in which its activities take place.

The evaluation process lasted three months, during which the initial data and post-commissioning measurements of the technological solution were collected. The interval was selected to ensure the operational stability of the fleet and reduce the influence of external variations such as high seasons or changes in transport demand.

Although the pre-experimental design allowed direct comparison of operational indicators before and after the intervention, the absence of a concurrent control group limits strict causal attribution. Therefore, the observed improvements cannot be interpreted as being caused exclusively by the Digital Twin system, since Hawthorne effects, short-term demand fluctuations, or other operational changes during the three-month observation period may also have influenced the results.

The population consisted of the 28 active vehicles of the evaluated company. A total of 25 vehicles were selected through non-probabilistic convenience sampling because they had fully functional OBD-II sensors and availability for continuous monitoring. Although a probabilistic method was not used, the sample represents 89.2% of the total fleet, providing high representativeness for the analysis.

The selection was justified by operational criteria: vehicles in continuous use, stable routes, and engines compatible with standard telemetry. The selection criteria made it possible to collect consistent and comparable data across both phases of the study.

B. Digital Twin Architecture

The architecture developed for the study was designed with a lightweight, scalable approach, leveraging low-cost technologies and open-source tools with high interoperability. The architecture is composed of four functional layers:

- IoT Layer (Data Capture): Integrated by OBD-II sensors installed in each vehicle, responsible for recording operational telemetry such as speed, acceleration, engine temperature, fuel consumption, and braking patterns. These data were transmitted wirelessly to the central server.
- Processing Layer (Microservices): Implemented in Python using the Flask framework. It included modules for data cleaning, outlier detection, unusual event identification, and daily indicator aggregation.
- Storage Layer (Database): MongoDB was chosen for its ability to manage both structured and unstructured data and for its flexibility in handling large volumes of telemetry. The database stored historical records, processed indicators, and instantaneous fleet statuses.
- Visualization Layer (Real-Time Dashboard): Developed to display processed information through graphs, alerts, and operational indicators. This layer simulated the fleet's behavior, enabling monitoring of trends, routes, critical events, and variations in load capacity.

Fig. 1 shows that the layers comprise the functional version of the Digital Twin, capable of synchronizing the physical and virtual states of the vehicles during their real operation.

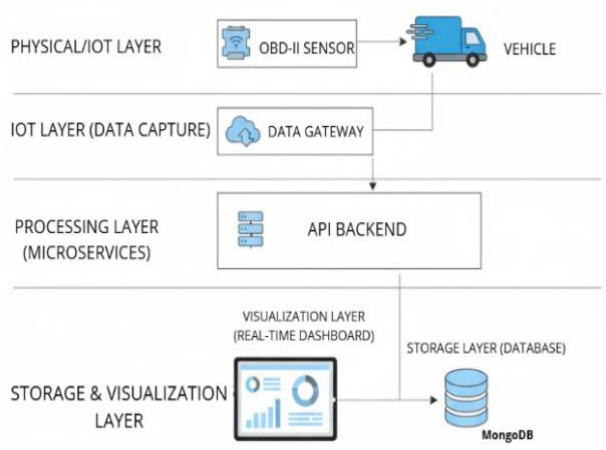


Fig. 1. Architecture of the lightweight Digital Twin system integrating OBD-II telemetry, microservice-based processing, a document-oriented database, and a real-time monitoring dashboard.

C. Procedure for Data Collection

The data collection process followed the following stages:

- Installation and calibration of OBD-II sensors in the 25 selected vehicles.
- Continuous telemetry record stored every 2 seconds throughout the study period.
- Data integrity validation, discarding corrupted or inconsistent values.
- Raw data processing, generating daily indicators per vehicle.
- Consolidation of operational indicators: frequency of accidents, load transported, and number of trips made.

The general data collection and processing workflow is presented in Fig. 2, which summarizes the sequence from sensor installation and telemetry acquisition to data validation, indicator generation, and operational analysis.

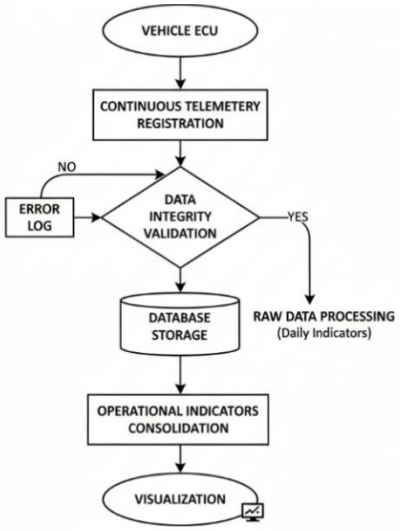


Fig. 2. Flow chart and processes.

The study focused on three key indicators:

- Accident frequency: number of incidents recorded per vehicle during the evaluated period.
- Load capacity use: average number of units transported per trip.
- Monthly number of trips: total trips needed to meet standard demand.

These indicators were selected for their direct relevance to efficiency, safety, and logistics performance.

D. Event Detection Module

The event detection module is responsible for identifying risky driving behaviors based on the telemetry received. It processes each new sample from the OBD-II device and evaluates a set of thresholds for speed, acceleration, and braking intensity. When one or more thresholds are exceeded, the module generates an event log that is stored in the database and displayed on the control panel as a security alert. Algorithm 1 summarizes the logic implemented in the event detection microservice.

Algorithm 1: Event Detection Module

```
Input: telemetry sample s = {speed, rpm, throttle, brakeForce, gpsPosition, timestamp}
Parameters:
SPEED_LIMIT = 90 km/h
ACC_THRESHOLD = 3.0 m/s² // harsh acceleration threshold
DECEL_THRESHOLD = 3.5 m/s² // harsh braking threshold
TEMP_LIMIT = 95 °C // engine temperature alert threshold
Output: list of detected events E
Begin
E ← ∅
//Overspeeding
    if s.speed > SPEED_LIMIT THEN
        E ← E ∪ {("OVERSPEED", s.timestamp, s.gpsPosition)}
    End if
// Harsh acceleration
    if s.acceleration > ACC_THRESHOLD then
        E ← E ∪ {("HARSH_ACCELERATION", s.timestamp, s.gpsPosition)}
    end if
// Harsh braking
    if s.brakeForce < −DECEL_THRESHOLD then
        E ← E ∪ {("HARSH_BRAKING", s.timestamp, s.gpsPosition)}
    end if
//Engine temperature alert
    if s.engineTemp > TEMP_LIMIT then
        E ← E ∪ {("HIGH_ENGINE_TEMPERATURE", s.timestamp, s.gpsPosition)}
    end if
    return E
    End
```

For the monitored fleet, the thresholds were set at 3.0 m/s² for harsh acceleration, 3.5 m/s² for harsh braking, and 95 °C for engine temperature alerts during the calibration stage of the system.

TABLE I. SYSTEM AND DATA ACQUISITION PARAMETERS

Parameters	Value
Number of vehicles monitored	25
Sensor Type	OBD-II (ELM327-based)
Telemetry frequency	1 sample/2 seconds
Data transmitted	Speed, RPM, throttle %, fuel rate, engine load, braking intensity
Transmission protocol	Bluetooth / WiFi
Database system	MongoDB (document-oriented)
Processing pipeline	Python microservices
Monitoring dashboard	Web-based (real-time)

Table I clearly shows how the system supporting the Digital Twin was configured. The company worked with a fleet of 25 units, each equipped with an OBD-II sensor capable of sending data at a short periodicity, sufficient to track vehicle behavior in near real-time.

The values collected provide a relatively complete view of each unit's state, ranging from basic variables such as speed to indicators of engine effort and the driver's brake use. All that information was transmitted wirelessly to a database designed to handle large volumes of records, where small Python modules then processed it.

The entire flow was integrated into a visualization panel that allowed for the review of the fleet's activity while it was in operation.

E. Monitoring Dashboard

In addition to data acquisition and processing, the proposed system includes a web dashboard that provides real-time visibility into fleet operations. The dashboard is organized into three main panels: a map view showing the current GPS position of the selected vehicle, a telemetry panel showing key indicators such as speed, engine temperature, and battery voltage, and an event panel listing active safety alerts.

Supervisors can filter events by vehicle, adjust the refresh rate, and toggle between the live and historical views. The interface allows operating personnel to monitor driving conditions during trips and review past events for training or audit purposes.

Fig. 3 shows a capture of the panel used to monitor the fleet in real time. There, you can see the vehicle's location, the central telemetry values, and the alerts generated during the operation.

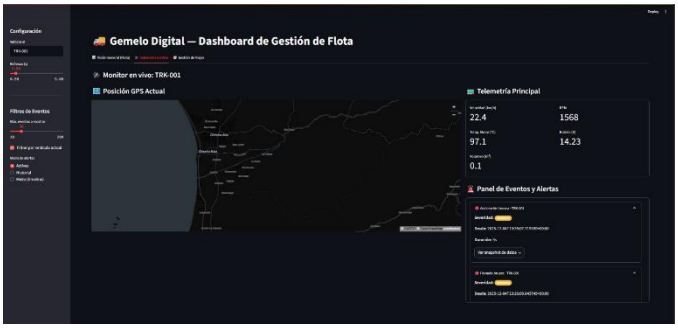


Fig. 3. Digital Twin–based fleet monitoring dashboard showing live GPS position, key telemetry indicators, and active safety alerts.

F. Statistical Analysis

T-tests were applied for related samples to compare the pre-test and post-test values of the three operational indicators. The significance level was set to $\alpha = 0.05$. The test was adequate given the continuous nature of the data and the comparison between two observation times for the same sample.

Likewise, descriptive measures such as averages, standard deviations, and percentage variations were used to complement the interpretation of the Digital Twin's impact. Additionally, basic assumptions, such as the normality of the differences, were verified through graphical and statistical analyses.

For the study, no routes, loads, or activities that could affect driver safety or the integrity of vehicle units were manipulated.

Regarding internal validity, efforts were made to control factors that could influence the results, such as changes in demand, route adjustments, or changes in fleet composition. Regarding external validity, it is recognized as limited; however, the procedure followed can be reproduced in companies operating under similar conditions, opening the possibility of applying the methodology in other contexts within the same sector.

V. Results

The incorporation of Digital Twin enabled unambiguous identification of changes in the fleet's performance throughout the analysis period. The comparison between the data obtained before and after implementation provides an accurate view of how the tool influenced three central aspects of the operation: safety, load capacity utilization, and the number of displacements required to meet demand. The most important findings are detailed below.

A. Frequency of Accidents

Before the implementation of Digital Twin, the company recorded minor accidents that, although they did not always lead to material damage, did indicate potential risks. On average, the fleet had 0.08 accidents per vehicle, which indicated driving behaviors requiring attention, such as sudden braking, unexpected maneuvers, or deviations from the planned route.

After the implementation of the IT solution, the indicator improved significantly, with no accidents recorded during the post-test period. Three main changes introduced through the Digital Twin system help explain the improvement. First, continuous monitoring of each vehicle made it possible to detect unusual behavior early and respond before it escalated into an incident. Second, the alerts sent to supervisory staff created opportunities for immediate corrective action during operation. Third, drivers tended to behave more cautiously because the most relevant aspects of their driving were being monitored in real time.

The statistical analysis indicated a significant association between the intervention period and the observed improvement in fleet management indicators.

B. Use of Load Capacity

In the initial stage, the average of the indicator was 31.04 units, reflecting variations in the allocation of goods and a tendency to send vehicles with available space, especially on routes with fluctuating demand.

After implementing Digital Twin, the average increased to 35.00 units, representing a significant improvement for the operation. The variation was not due to an increase in demand, but to a more innovative use of information. The system allowed:

- Identify trips with low occupancy.
- Reorganize load allocation more evenly.
- Avoid unnecessary departures when consolidating goods is possible.

One of the most apparent effects was that supervisors could visualize each vehicle's occupancy in real time, facilitating more precise decisions when coordinating dispatches. In practice, the new monitoring process helped prevent a routine pattern in the company: sending vehicles with low loads just to meet internal schedules, even when it was not strictly necessary.

C. Monthly Number of Trips

A relevant aspect of logistics management is the number of displacements made to meet a constant demand. During the pre-test period, the number of trips tended to be higher than necessary, in part because decisions were made with incomplete information or based on estimates made at the start of each day.

Among the improvements highlighted by the operational team are:

- Fewer "empty" or underutilized trips.
- Better allocation of units by load type.
- Greater anticipation of delays or detours enabled reorganizing routes without generating additional trips.

The statistical analysis showed significant differences between the pre- and post-intervention stages, supporting the conclusion that the Digital Twin improved efficiency, reduced unnecessary travel, and contributed to lower operating costs for fuel, maintenance, and person-hours.

All three indicators point in the same direction: the Digital Twin had a positive, measurable impact on fleet operations. Not only did it improve safety and productivity, but it also offered a stronger basis for decision-making.

The results obtained show that:

- Accidents were eliminated during the post-test period.
- Use of load capacity was elevated, reducing underutilization.
- The number of trips was reduced, reflecting a more rational use of resources.

Therefore, the changes observed are consistent with the use of more accurate and timely information provided by the Digital Twin architecture. However, because the study used a pre-experimental design without a concurrent control group, these improvements should not be interpreted as being caused exclusively by the intervention. Table II shows the comparison between the pre- and post-test results.

TABLE II. PRE-TEST AND POST-TEST OPERATIONAL INDICATORS

Indicators	Pre-test and post-test operational indicators		
	Pre-test	*Post-test*	*Change*
Accident frequency	0.08	0.00	-0.08
Load capacity utilization (%)	31.04	35.00	3.96
Monthly trip count	250	230	-20

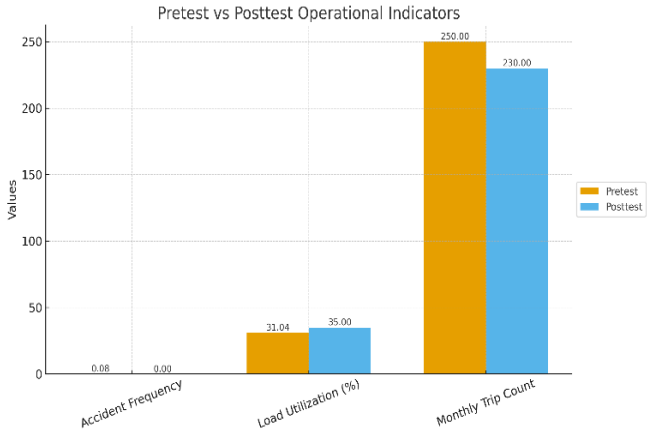


Fig. 4. Pre-test and post-test operational indicators.

Fig. 4 shows how the three operational indicators changed after implementing the Digital Twin. We can see a drop in incidents, better use of capacity, and fewer trips needed to meet the same demand.

In short, experience shows that even a company with limited resources can obtain significant benefits by adopting technologies that integrate telemetry, data analysis, and operational visualization.

VI. DISCUSSION

The results obtained allow a clearer understanding of the real impact that a Digital Twin system can have when integrated into an operating environment such as that of a freight transport company. In the first instance, the total reduction in accident frequency (from an average of 0.08 to 0.00) confirms that technology serves not only as a monitoring mechanism but also as an instrument that changes behavior and reduces risk exposure. The reduction also helped the company avoid repair costs and vehicle downtime. According to internal estimates from the head of the operations area, even a minor incident of this kind can lead to one or two weeks of service interruption for a single vehicle, in addition to direct maintenance expenses. Such findings are consistent with reports on DT-ITS systems, which emphasize the ability of Digital Twins to improve interoperability between devices, anticipate incidents, and enhance road safety through continuous synchronization between physical and virtual environments [11].

Likewise, the significant improvement in load capacity (increased from 31.04 to 35.00 units) demonstrates that the availability of reliable, real-time data enables reorganizing assignments, reducing underutilized trips, and optimizing vehicle use. The result is consistent with the findings of Moshood et al. [12], who argue that Digital Twins enhance logistical visibility, detect structural inefficiencies, and help manage critical resources within a sustainability framework. The company stopped operating on estimates and began basing its decisions on evidence.

Regarding the decrease in the number of trips (250 trips to 230), although the indicator did not decrease drastically, the difference was significant ($p = .045$). That is, as the load capacity used per trip increases, the need for additional travel is automatically reduced. Wang et al. demonstrated that decision-making based on real-time synchronized information optimizes routes, avoids redundant displacements, and improves operational efficiency in complex environments [13].

The combination of higher cargo utilization and fewer monthly trips also evidences improvements in fuel efficiency and labor use. Considering a conservative fuel cost and average driving time per route, the reduction translates into savings in fuel consumption, driving hours, and overtime pay. The implementation of Digital Twins contributes not only to safer operations but also to a gradual reduction in operating costs.

Discussion of the results is also related to experiences reported in other countries. Tang et al. [14], for example, demonstrated that combining IoT with predictive maintenance strategies reduces downtime and prevents operational failures. Similarly, Ramasubramanian et al. [15] argued that Digital Twins support strategic decision-making related to fleet

management. Although both studies were carried out in more technologically advanced settings, the results of the present study suggest that Digital Twins can also deliver significant benefits in companies with limited resources.

The results show that Digital Twins are a feasible alternative for SMEs. Even without having a sophisticated technological infrastructure, it was possible to propose concrete improvements in security, performance, and resource optimization. Implementing Digital Twins can contribute not only to safer operations but also to a gradual reduction in operating costs.

The research was conducted in an SME in the freight transport sector, so the findings cannot be generalized to larger organizations with more complex processes and infrastructure. As Rigó et al. [5] point out, the performance of a Digital Twin is closely linked to the degree of technological maturity and the institutional commitment to adopting Industry 4.0 tools. These factors can vary considerably across companies.

Several factors conditioned the scope of the study. First, we worked with a single company and a small number of units, which limits the ability to extend the results to other environments with different characteristics. In addition, the implemented solution lacked more sophisticated predictive models, making it impossible to analyze long-term behavior or trends. In addition, differences in infrastructure and level of digitization between companies can influence the degree of replicability of the proposal.

VII. Conclusion

The findings confirm that the incorporation of a Digital Twin system had a positive, clearly observable effect on the management of the evaluated cargo fleet. The data obtained show improvements in three fundamental aspects of the operation: safety, use of load capacity, and travel efficiency. The absence of accidents during the post-test is a clear sign of the value of continuous, real-time monitoring as a tool for anticipating contingencies. Likewise, the improvement in the use of available capacity suggests that the IT solution implemented supported more accurate planning and more efficient use of resources.

A moderate reduction in the number of trips indicates that vehicle units will make fewer trips; as a result, operating costs will be reduced. The finding is consistent with prior studies that highlight the role of Digital Twins in optimizing logistics processes through the timely availability of information.

The results show that the implementation of Digital Twin technology enabled a transition from a traditional, transaction-based environment to a more data-driven setting that better supports operational decision-making. It is also confirmed that deploying an accessible solution developed with open-source tools delivers significant improvements for SMEs.

VIII. Future Work

According to the findings and limitations of the study, several lines of research are identified that could strengthen the use of Digital Twin technology in fleet management:

- Integrating machine learning with Digital Twin technology can reduce empty trips and operational failures by improving prediction and operational planning [14], [18].
- Extension of the architecture towards a complete Digital Supply Chain Twin. The studies by Gerlach et al. [18], Hakimi et al. [19] and Fatorachian et al. [20] underscore the importance of scaling Digital Twins across the entire logistics chain, not just the fleet.
- Advanced simulation of urban scenarios. Using ROS2 or AWSIM platforms would enable us to model traffic behavior and evaluate more efficient routes in real time.
- Integration with TMS and ERP systems. Connecting the DT with corporate systems would make it easier to automate load assignments, validations, reports, and prioritizations [21].

References

[1] D. Wu *et al*., "Digital Twin Technology in Transportation Infrastructure: A Comprehensive Survey of Current Applications, Challenges, and Future Directions," Feb. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app15041911.

[2] D. Nag, F. Brandel-Tanis, Z. A. Pramestri, K. Pitera, and Y. K. Frøyen, "Exploring digital twins for transport planning: a review," Dec. 01, 2025, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1186/s12544-025-00713-0.

[3] E. Markopoulos *et al*., "The Impact of Digital Twins Technology in Maritime Fleet and Safety Management," in *Accessibility, Assistive Technology and Digital Environments*, AHFE International, 2024. doi: 10.54941/ahfe1004619.

[4] L. J. Castillo Rabanal, A. S. Huamanchumo Gordillo, and M. A. Lecca Rengifo, "The Impact of the Application of Digital Twin Technology in Supply Chain 4.0: A Systematic Review," *Gestión de Operaciones Industriales*, vol. 3, no. 1, pp. 80–92, Sep. 2024, doi: 10.17268/goi4.0.2024.05.

[5] L. Rigó, J. Fabianová, M. Lokšík, and N. Mikušová, "Utilising Digital Twins to Bolster the Sustainability of Logistics Processes in Industry 4.0," *Sustainability (Switzerland)* , vol. 16, no. 6, Mar. 2024, doi: 10.3390/su16062575.

[6] U. Asad, M. Khan, A. Khalid, and W. A. Lughmani, "Human-Centric Digital Twins in Industry: A Comprehensive Review of Enabling Technologies and Implementation Strategies," Apr. 01, 2023, *MDPI*. doi: 10.3390/s23083938.

[7] J. F. Yao, Y. Yang, X. C. Wang, and X. P. Zhang, "Systematic review of digital twin technology and applications," Dec. 01, 2023, *Springer*. doi: 10.1186/s42492-023-00137-4.

[8] W. Li, H. Zhou, Z. Lu, and S. Kamarthi, "Navigating the Evolution of Digital Twins Research through Keyword Co-Occurence Network Analysis," Feb. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/s24041202.

[9] I. F. Iumanova, P. V. Matrenin, and A. I. Khalyasmaa, "Review of Existing Tools for Software Implementation of Digital Twins in the Power Industry," Oct. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/inventions9050101.

[10] S. Werbińska-Wojciechowska, R. Giel, and K. Winiarska, "Digital Twin Approach for Operation and Maintenance of Transportation System—Systematic Review," Sep. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/s24186069.

[11] C. Ge and S. Qin, "Digital twin intelligent transportation system (DT-ITS)—A systematic review," Dec. 01, 2024, *John Wiley and Sons Inc*. doi: 10.1049/itr2.12539.

[12] T. D. Moshood, G. Nawanir, S. Sorooshian, and O. Okfalisa, "Digital twins driven supply chain visibility within logistics: A new paradigm for future logistics," Jun. 01, 2021, *MDPI AG*. doi: 10.3390/asi4020029.

[13] K. Wang *et al*., "Smart Mobility Digital Twin Based Automated Vehicle Navigation System: A Proof of Concept," *IEEE Transactions on*

Intelligent Vehicles, vol. 9, no. 3, pp. 4348–4361, Mar. 2024, doi: 10.1109/TIV.2024.3368109.

[14] Z. Tang, D. Zhuang, and J. Zhang, "Evaluation framework for domain-specific digital twin platforms," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-024-82154-8.

[15] A. K. Ramasubramanian, R. Mathew, M. Kelly, V. Hargaden, and N. Papakostas, "Digital Twin for Human-Robot Collaboration in Manufacturing: Review and Outlook," *Applied Sciences (Switzerland)*, vol. 12, no. 10, May 2022, doi: 10.3390/app12104811.

[16] I. Abdullahi, H. Larijani, D. Liarokapis, J. Paterson, D. Jones, and S. Murray, "A Data-Intelligence-Driven Digital Twin Framework for Improving Sustainability in Logistics," *Applied Sciences (Switzerland)*, vol. 15, no. 2, Jan. 2025, doi: 10.3390/app15020601.

[17] N. S. Madusanka, Y. Fan, S. Yang, and X. Xiang, "Digital Twin in the Maritime Domain: A Review and Emerging Trends," May 01, 2023, *MDPI*. doi: 10.3390/jmse11051021.

[18] B. Gerlach, S. Zarnitz, B. Nitsche, and F. Straube, "Digital Supply Chain Twins—Conceptual Clarification, Use Cases and Benefits," *Logistics*, vol. 5, no. 4, Dec. 2021, doi: 10.3390/logistics5040086.

[19] F. Hakimi, T. Khaled, M. Al-Kharaz, A. C. F. Gouabou, and K. Amzil, "Towards a Digital Twin Modeling Method for Container Terminal Port," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 3113–3121. doi: 10.1016/j.procs.2024.09.361.

[20] H. Fatorachian, H. Kazemi, and K. Pawar, "Enhancing Smart City Logistics Through IoT-Enabled Predictive Analytics: A Digital Twin and Cybernetic Feedback Approach," *Smart Cities*, vol. 8, no. 2, Apr. 2025, doi: 10.3390/smartcities8020056.

[21] A. Dorofeev, V. Kurganov, N. Filippova, A. Petrov, D. Zakharov, and S. Iarkov, "Improving Transportation Management Systems (TMSs) Based on the Concept of Digital Twins of an Organization," *Applied Sciences (Switzerland)*, vol. 14, no. 4, Feb. 2024, doi: 10.3390/app14041330.

Real-Time Thought-to-Vision Generation Using Low-Channel EEG and Feature-Fusion Learning

Abha Marathe, Medha Wyawahare, Milind Rane, Vrinda Parkhi,
Milind Kamble, Anvita Barlingay, Anushka Goswami, Aditya Awate
Department of Electronics and Telecommunication Engineering, Vishwakarma Institute of Technology, Pune, India

***Abstract*—Severe motor disabilities and paralysis make it hard for individuals to communicate their thoughts and express their imagination using standard interfaces. Recent methods that convert EEG signals into images, using diffusion models, have shown superior results. However, these methods usually depend on high-density EEG systems with 32 to 128 channels, deep neural EEG encoders, and large datasets. This leads to high computational costs, poor real-time performance, and limits their use in assistive settings. To address these problems, this paper proposes a Thought-to-Vision system that is lightweight and real-time. In this work, Thought refers specifically to the imagination of simple geometric shapes (circle, square, and triangle) under controlled experimental conditions. This Thought-to-Vision system can decode the imagined geometric shapes from a low-channel EEG system that only requires 2 channels and then produce visual images based on a diffusion model. The EEG signal was recorded at 250 Hz with 150 trials per session, consisting of 50 trials for each circle, square, and triangle shape. The signal was filtered using artifact rejection, 50 Hz notch filtering, and bandpass filtering between 1 and 40 Hz. A Tri-Domain EEG feature fusion (TDEF) that combines spectral features (FFT band power), Time-Frequency features (Daubechies-4 wavelet coefficients), and statistical features was developed and tested against several benchmarks. These included feedforward networks, CNNs, LSTM/GRU-based time-series encoders, CNN-Transformer models, and EEG-CLIP alignment. Evaluation is measured using classification accuracy, precision, recall, and F1 score, along with embedding consistency for semantic alignment. The experimental results indicate that the TDEF with the XGBoost classifier reaches around 94% for classification accuracy, precision, recall, and F1-score. This performance surpasses deep time-series encoders, which achieved up to 39.09% accuracy, and contrastive EEG-CLIP models, which had 82.97% accuracy. The classified EEG embeddings were then used to guide a latent diffusion model, enabling coherent and semantically consistent image generation. These findings confirm that feature-fusion learning with XGBoost can outperform deep EEG encoders in low-channel situations. This offers a solid, efficient, and practical solution for real-time assistive brain-computer interfaces.**

Keywords*—Brain computer interface; feature fusion learning; diffusion models; EEG to image generation; wavelet; EEG signal processing

I. Introduction

The demodulation of human thought and imagination patterns via neural activity is one of the ever-green areas in brain-computer interface research [1]. Electroencephalography (EEG), due to its non-invasive nature and high temporal resolution, has been investigated for its use in cognitive processing translation into machine-readable signals for several applications [2]. While fMRI is characterized by its ability to produce highly accurate images, it is associated with poor temporal resolution and excessive costs. On the other hand, EEG, though having poor spatial resolution, has excellent temporal resolution as well as portability. However, because of its poor spatial resolution, it can only be used for the interpretation of crude representations of visual images, not their minute details. The translation of imagination-related EEG patterns into a meaningful pictorial interpretation has a wide scope in assistive communication systems, neurofeedback, and human-machine interaction contexts. Recent works have focused on the usage of deep neural networks and diffusion-based generative models for the reconstruction of images from brain signals using EEG-to-image synthesis. The potential of aligning the embedding spaces of the brain signals with the visual embedding space has been shown using frameworks such as Brain-Dreamer [3], Dream-Diffusion [4], and EEG-Vision [5]. These frameworks require high-density EEG recording (32 to 128 channels), a massive dataset, and computationally expensive brain decoders, thus making it very difficult to implement them in a realistic scenario. Yet, there is an essential gap in research work: the existence of light-weight EEG-to-image models able to realize real-time processes on a low-channel EEG signal with a high degree of accuracy. In fact, when it comes to actual applications involving a person with motor disorders, it can be more essential to ensure simplicity and portability rather than the generative capability. The over-reliance on a deep encoder when it comes to EEG often results in a poor generalization ability. With the shortcomings in mind, the proposed research seeks to develop a low-channel, real-time Thought-to-Vision generation framework where simplicity of usability is put foremost. Contrary to the requirement of using deep EEG encoders in the existing work, the proposed approach focuses on the concept of feature fusion-based learning, which exploits multiple complementary descriptors of the EEG signal to increase discriminatory efficiency without the associated heavy computations. The key contributions of the current study are the design of a light-weight, two-channel EEG-based Thought-to-Vision system for real-time applications, and the development of a tri-fusion EEG feature representation scheme that combines spectral, time-frequency, and statistical features to efficiently capture complementary information from the brain signals. A feature-fusion–based XGBoost classifier is employed, which consistently outperforms state-of-the-art deep EEG encoders in low-channel settings. Furthermore, the system achieves efficient EEG decoding coupled with diffusion-based

image generation, allowing the visualization of imagined content in a coherent and semantically meaningful manner. Finally, a comprehensive comparative evaluation of classical, feature-based, and deep learning EEG encoders is presented, clearly demonstrating the advantages of feature-based learning approaches for robust, lightweight, and assistive brain–computer interface applications.

This work makes four key contributions: (1) a real-time, low-channel (2-channel) EEG-based Thought-to-Vision framework suitable for assistive deployment, (2) a tri-fusion EEG feature representation combining spectral, time–frequency, and statistical descriptors, (3) a comprehensive empirical comparison of feature-based and deep time-series EEG encoders under low-channel constraints, and (4) a diffusion-conditioned image generation pipeline driven by lightweight EEG decoding.

The rest of the document is structured as follows: Literature review is presented in Section II followed by Methodology in Section III. Experimental results in Section IV, and Conclusion and Future Scope in Section V.

II. Literature Review

The literature review is categorized into five major groups: (1) EEG Signal Processing and Fundamental Concepts, (2) EEG to Image Generation using Deep Learning and Diffusion Models, (3) Multimodal Learning, Retrieval, and Representation Learning, (4) Applications of EEG in Creative, Medical, and Assistive Systems, (5) Surveys, General BCI Systems, and Recent Trends.

A. EEG Signal Processing and Fundamental Concepts

Early research by Kumar and Bhavaneswari [2] laid the foundation for EEG signal analysis by categorizing EEG activity into Delta, Theta, Alpha, Beta, and Gamma frequency bands using Fourier and wavelet-based spectral techniques. Their work demonstrated that band-power features, particularly Alpha and Beta rhythms, are reliable indicators of cognitive and mental states, forming the basis for later EEG preprocessing and feature extraction pipelines. In addition to model development, Hatton et al. [6] investigated various EEG acquisition systems and proved that multi-channel systems can greatly enhance spatial and spectral representation of EEG signals, providing valuable guidance for EEG system implementation.

B. EEG-to-Image Generation Using Deep Learning and Diffusion Models

With the rise of deep learning, EEG-based visual decoding advanced significantly. Wang et al. [3] proposed BrainDreamer, a language guided EEG to image generation framework that aligns EEG, text, and image embeddings in CLIP latent space using contrastive learning and diffusion-based image synthesis. This method facilitated semantically coherent and controllable image reconstruction from EEG signals.Bai et al. [4] proposed Dream Diffusion, which eliminated the need for text-based guidance and proved that diffusion models are superior to GAN based methods for EEG to image reconstruction, achieving more realistic visual reconstructions. Huangtao et al. [5] introduced EEG-Vision, which shows integration of time-frequency EEG features with diffusion models to achieve high quality visual reconstructions. To enhance semantic coherence, Yang and Liu [7] proposed a method integrating CNN based EEG classification with diffusion models, enhances both decoding performance and visual coherence. Chen et al. [8] investigated diffusion-based EEG image reconstruction and proved that structured EEG encoding can enhance reconstruction performance. Similarly, Li et al. [9] proposed a guided diffusion model that utilized EEG embeddings to improve reconstruction performance. Lopez et al. [10] proposed a latent diffusion-based EEG to image generation framework using simulated EEG data, demonstrating the potential for training generative models at scale.

C. Multimodal Learning, Retrieval, and Representation Learning

Retrieval based methods further extended the field of EEG vision research. Sithisint et al. [11] proposed EEG2Face Query, a method for retrieving facial images from EEG signals via CNN and Vision Transformer based latent representation learning. Zhang et al. [12] proposed Cognition Capturer, which associated EEG with multimodal embeddings (image and text) to improve visual decoding performance. Ferrante et al. [13] demonstrated that knowledge distillation enables compact EEG encoders to retain performance when paired with latent diffusion models. Puah et al. [14] proposed a model based on a diffusion model framework, "EEG-DGM" or "EEG Diffusion Model" suggests a novel approach to learning representations of EEG signals using a denoising and diffusion probabilistic model which would be beneficial to the field of EEG classification. Large scale benchmarking was enabled by Zhu et al. [15], who introduced EEG ImageNet, a dataset with multigranularity labels for standardized evaluation. Pan et al. [16] proposed a deep visual representation model for reconstructing stimulus images from EEG signals. Spampinato et al. [17] proposed a deep learning-based model using recurrent neural networks for discriminative embedding learning using EEGs. It is notable in the sense that their work emphasizes the importance of temporal information rather than spatial data since, even though fMRI can capture higher spatial data while being expensive, it lacks the temporal aspect. While the alternative MEG has the advantage of improved spatial resolution, it is non-portable and difficult to operate compared to other modalities such as EEG. This proves that the need for good representation learning through effective embeddings is a necessity for visual decoding from EEG signals. Finally, Zhang et al. [18] extended Cognition Capturer in a journal version, reinforcing the importance of multimodal alignment for high-quality EEG-based visual decoding.

D. Applications of EEG in Creative, Medical, and Assistive Systems

Earlier creative BCI applications like Brain Painting by Münßinger et al. [19] demonstrated the viability of using EEG for creative tasks, particularly for ALS patients. In addition to 2D reconstruction, Guo et al. [20] introduced Neuro3D, demonstrating that it is possible to reconstruct 3D objects from EEG signals using diffusion-based point cloud generation. The application of EEG technology also made its way into creative applications. Liu and Wang [21] proposed Mental Gen, an EEG controlled generative model, for interior space design, which showed the possibility of using brain signals for personalized creative tasks. Soroush et al. [22] systematically reviewed the

EEG characteristic and through experimental design highlighted the EEG driven creativity research. Medical EEG applications were discussed by Foreman et al. [23], employing the quantitative EEG characteristics for the detection of brain ischemia. Pujac et al. [24], shown the artistic applications by visualizing real time EEG data through generative AI art.

E. Surveys, General BCI Systems, and Recent Trends

More general views on EEG based BCIs were presented by Lazarou et al. [1], reviewing the EEG based communication and rehabilitation systems, with a focus on robustness and clinical applicability. A survey conducted by Shukla et al. [25] investigated the combination of EEG and generative AI, noting the importance of diffusion models and multimodal learning. EEG based creativity research was further reviewed by Zeng and Soroush [26], while Liu et al. [27] surveyed recent BCI applications. Cognitive inspired reasoning models like Sketch of Thought were presented by Aytes et al. [28], who proposed light weight representations for efficient reasoning. Advanced cognitive efficiency was revisited by Aytes et al. [29], extending Sketch-of-Thought for adaptive reasoning. In addition, Majima and Nishimoto [30] investigated reconstructing visual imagery from brain activity by applying Bayesian estimation through deep neural networks. This research shows the ability to reconstruct visual imagery from neural activity, thereby reinforcing the possibility of mapping neural activities to visual imagery.

III. METHODOLOGY

The methodology of the proposed work is shown in Fig. 1. It contains the following steps:

A. Dataset Collection

A raw, synchronized dataset of 2-channel EEG signals recorded while a human participant views a series of images shown in Fig.2. The main aim was to record the brain's electrical signals associated with the visual perception of geometric shapes. EEG signals were recorded using a low-cost two-channel recording system. Data was collected using a controlled visual stimulation paradigm to record consistent neural responses associated with basic geometric shape perception. During the recording process, three shapes were presented before the subject i.e., circle, square, and triangle, with 50 different stimuli for each shape, making a total of 150 trials per recording session. Each stimulus was displayed for 500ms, followed by a 500ms inter-stimulus interval using a blank screen to avoid overlap of successive cognitive responses. To avoid fatigue and cognitive drift, trials were arranged in shape blocks with short rest intervals between shape blocks. EEG signals were recorded at a sampling rate of 250 Hz and synchronized with the stimulus display using event marker logs, allowing for precise segmentation of neural responses associated with each visual stimulus.

B. Data Processing

To increase the quality of the signal and guarantee the extraction of the unique features of the neural activity, a preprocessing step was carried out on the raw EEG data. The EEG signals were recorded at a frequency of 250 Hz, and trial-specific segments were extracted around the synchronized event markers. The preprocessing pipeline included three steps in a sequence: Artifact rejection, removal of power line noise, and band pass filtering. Each of these steps was designed to remove non neural interference while retaining the cognitive related brain signals.

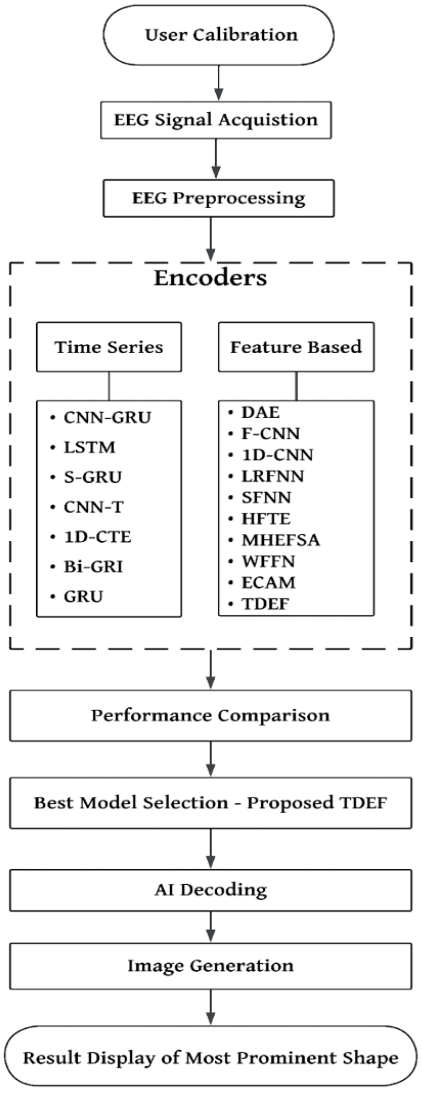


Fig. 1. Flowchart of the project.



Fig. 2. Dataset collection process.

First, we discarded the epochs that were artifact rich by using an amplitude-based rejection, which was focused on non-cortical physiological artifacts such as eye blinks and muscle activity. Any trial that had instantaneous EEG amplitudes outsides the ±100 μV range was rejected, as is standard practice in ERP analysis. In formal terms, any trial i, with EEG signal is rejected for the condition given in (1).

$$\exists t : |x_i(t)| > \theta_{artifact} \quad (1)$$

The value of $\theta_{artifact} = 100.0\mu V$.

This criterion is extremely helpful in suppressing the large artifact perks without affecting the trials that show the existence of neural activity.

Secondly, to eliminate the power line noise we used a notch filter at 50 Hz. It is an IIR notch filter with Q=30. It is used as a zero- phase forward and backward filter to eliminate any phase shift. This ensures that the temporal fidelity of the neural oscillations is maintained and the power line interference is removed with absolute accuracy.

Finally, the EEG signals are band pass filtered between 1 to 40 Hz using a fifth order Butterworth filter to zero in on the frequency of interest for cognitive processing. The squared magnitude response of the filter is given (2).

$$|H(j\omega)|^2 = \frac{1}{1+(\omega/\omega_c)^{2N}} \quad (2)$$

In this setup, N=5 signifies the order of the filter, while ω = 2πf and the cutoff ωc is varied to 1 Hz and 40 Hz. This filtering step eliminates the slow drift of the baseline and noise but eliminates the EEG cycles: Delta (1-4 Hz), Theta (4-8 Hz), Alpha (8-13 Hz), Beta (13-30 Hz), and low Gamma (30-40 Hz) cycles.

C. Feature Engineering

Conversion of feature space in EEGs to feature space in vision assumes that the neural activity carries enough information about the object seen or imagined. This hypothesis has been demonstrated by previous research studies such as the one performed by Spampinato et al. [29], in which machine learning techniques were utilized to learn discriminative feature embeddings from EEGs. They concluded that neural activity could be mapped to a structured feature space for visual tasks, thereby supporting the feasibility of EEG to latent space mapping in visual decoding frameworks.

The raw EEG signal was then converted into a clean, interpretable form that can be used for classification as well as image reconstruction. For each trial, we extract the salient features and convert them into unique spectral time frequency, and statistical patterns of the signals. We start by converting the signal from time to frequency using a fast Fourier transform, extracting power in the Delta, Theta, Alpha, Beta and Gamma frequencies. At the same time, we apply a Discrete Wavelet Transform with a Daubechies-4 wavelet to decompose the signal into multi-resolution sub-bands (A4, D4-D1), which helps to reveal the non-stationary and transient brain activity. However, since the raw wavelet coefficients are high dimensional and redundant, we reduce them by computing statistics: mean, standard deviation, and median of absolute values for each set of coefficients. This transforms time frequency data into compact, meaningful feature vectors. We then stack the spectral features with the wavelet statistics for both EEG channels into a unified feature vector. Finally, each trial is represented by a fixed 40-dimensional feature vector that retains the salient signal information while removing noise, variability, and computational complexity. While the suggested representation method represents every trial as a fixed- length vector, the fact that there are temporal relationships among EEG signals is acknowledged. In order to incorporate temporal aspects, sequence-based models such as LSTM, GRU, CNN-GRU, and transformer architectures are investigated within the framework of the experiment. The models are able to learn about the temporal characteristics of raw EEG signals and enable time-dependent analysis. Nonetheless, the results obtained from the experiments show that in low-channel configurations, temporal encoding alone cannot solve the problem since the signals are too similar. This strategy directly injects domain knowledge into the learning process, alleviating the need for data intensive deep encoders and improving channel limited and small sample robustness.

D. Encoder Based Algorithms

In order to compare the representation methods of EEG data using the constraint of a few channels, we focused on two broad categories of encoders: 1) feature-based encoders and 2) time-series-based encoders. This categorization allows us to compare domain knowledge-based learning with deep learning for representation.

1) Time series based encoders: Seven different time series encoders were experimented first. The details of the experimentation are discussed in the below section.

a) CNN-GRU hybrid encoder (CNN-GRU): A hybrid approach was also explored that combined one-dimensional convolutional neural networks (1D CNNs) with a Gated Recurrent Unit (GRU) to determine the effectiveness of temporal modeling as a subsequent step following the convolutional extraction of features. The model was trained directly on standardized raw two-channel EEG time-series data, using convolutional layers first to extract local temporal features, followed by a recurrent layer. The model consisted of two 1D convolutional layers, the first of which had 64 filters with a kernel size of 10, and the second of which had 128 filters with a kernel size of 10. After each convolutional layer, batch normalization and max pooling were used to reduce the temporal resolution and remove noise. The output feature vectors were then fed into a GRU layer with 128 hidden units. After this, the feature vector was projected into a 512-dimensional embedding space using fully connected layers, with L2 normalization. Training used triplet margin loss with $\alpha = 0.4$, and the Adam optimizer with a learning rate of 1×10^{-4}, batch size of 32, and early stopping with a maximum of 200 epochs. The model had a relatively low-test accuracy of 31.98%, and a confusion matrix that showed a strong collapse to a single predicted class with little distinction between the remaining classes. This indicated that the convolutional preprocessing alone was insufficient to model the low temporal

separability inherent in low-channel EEG signals when trained without feature-level or cross-modal supervision.

b) Long Short-Term Memory (LSTM): A pure Long Short-Term Memory (LSTM) encoder was tested to assess whether raw electroencephalography (EEG) time-series data could be modeled directly to derive a discriminative pattern of neural activity with a small number of channels. The LSTM encoder was trained on normalized raw EEG time-series data from two-channel recordings, where each trial was a fixed-length temporal sequence. Before training, time-series data were standardized per channel to improve the robustness of the optimization procedure. The encoder model consists of a batch normalization layer followed by a single LSTM layer with 128 hidden units, and the final hidden state is projected into a 512-dimensional L2-normalized embedding space through fully connected layers. Training was performed using triplet margin loss ($\alpha = 0.4$) and the Adam optimizer (learning rate = 1×10^{-4}) with a batch size of 32, including early stopping and learning rate scheduling, for a maximum of 200 epochs. Although the model was trained, it only reached a moderate accuracy of 33.50%, with a single imagined shape largely confused, as revealed by the confusion matrix. These results indicate that direct modeling of time series without considering spatial, spectral, or feature-level information is inadequate for accurate EEG decoding with a small number of channels.

c) Stacked GRU Network (S-GRU): The effectiveness of a stacked Gated Recurrent Unit (GRU) encoder was tested to evaluate the impact of hierarchical modeling of temporal patterns in improving discriminative learning of raw electroencephalography (EEG) time series data, especially when there is a constraint on the availability of channels. The stacked GRU encoder was trained on normalized two-channel EEG data, where each trial was represented as a fixed-size temporal array. Before training, the signals were normalized channel wise to facilitate stable optimization. The stacked encoder consists of two GRU layers with 128 hidden units each, where the first layer produces full temporal arrays, and the second layer produces only the final hidden state. This hierarchical representation is followed by fully connected layers and a projection into a 512-dimensional, L2 normalized embedding space. Training was performed using triplet margin loss ($\alpha = 0.4$) with the Adam optimizer at a learning rate of 1×10^{-4}, a batch size of 32, and early stopping with learning rate scheduling over a maximum of 200 epochs. The stacked GRU network reported an accuracy of 33.76%, and the confusion matrix revealed a strong class collapse to a single predicted class. These results suggest that simply adding depth to recurrent models is not sufficient to overcome the difficulties associated with temporal-only EEG decoding, especially when there is a constraint on channel access.

d) CNN–Transformer Encoder (CNN-T): A hybrid encoder combining both convolutional and transformer modules was tested to assess the role of self-attention mechanisms in the learning of temporal patterns in low-channel EEG signals when combined with convolutional feature extraction. The model was trained directly on standardized raw two-channel EEG time-series data, where convolutional layers were used to extract local temporal features, which were then followed by global modeling with attention mechanisms. In particular, two one-dimensional convolutional layers with 64 and 128 filters, respectively, with a kernel size of 10, were used, with max pooling to reduce the temporal dimensionality. The output feature maps were then fed into a transformer encoder layer with multi-head self-attention (four heads, each with an attention head dimension of 128), with residual connections, layer normalization, and a feed-forward network. Global average pooling was used to aggregate temporal information, followed by a fully connected projection layer that embedded the data into a 512-dimensional L2-normalized space. Training was done using triplet margin loss with $\alpha = 0.4$, optimized using the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 32, and early stopping with learning-rate scheduling over a maximum of 200 epochs. Despite the use of self-attention, the model resulted in an accuracy of 33.76%, and the confusion matrix showed a complete collapse to a single predicted class. These results suggest that when self-attention mechanisms in the transformer model are used on low-channel EEG time-series without feature engineering or cross-modal learning, it has a tendency to magnify dominant but non-discriminative temporal patterns rather than improve class discrimination.

e) 1D Convolutional Time-Series Encoder(1D-CTE): The effectiveness of one-dimensional convolutional time-series encoder (1D-CTE), a one-dimensional convolutional neural network (1D-CNN) model, was evaluated to assess the effectiveness of hierarchical temporal feature representation by itself for discriminative EEG decoding in channel-constrained settings. The model was trained on normalized raw two-channel EEG time-series data, where each trial was represented as a fixed-size temporal vector. Normalization was performed across channels before training to improve the robustness of optimization. The encoder neural network architecture is designed with three layers of 1D convolution with increasingly larger filter sizes of 64, 128, and 256, with a kernel size of 10, followed by max-pooling layers to enable hierarchical temporal feature representation at multiple scales. A global average pooling layer is used to obtain the temporal representation, and then a fully connected layer is used to map the features into a 512-dimensional embedding space, which is further L2-normalized. The model trains with a triplet margin loss function with alpha parameter value of 0.4, and it trains with the Adam optimizer with a learning rate of 1-4. The training is performed in batches of 32, with early stopping and learning rate scheduling, up to a maximum of 200 epochs. The accuracy achieved by the model was 35.03%, and the results obtained from the analysis of the confusion matrix showed that there was a highly dominant distribution towards a single class label with a large overlap between the class labels. The above results clearly indicate that hierarchical temporal feature representation alone, without any additional spectral or feature-level structure imposed through convolutional processing, is insufficient for accurate EEG decoding in channel-limited scenarios.

f) Bidirectional GRU (Bi-GRU): The bidirectional Gated Recurrent Unit (Bi-GRU) encoder was also employed to

determine whether the combination of both forward and backward information could help in the decoding of the EEG signal in the low-channel scenario. The training was performed on the normalized raw two-channel EEG time series data, where each trial was represented by a fixed temporal pattern. The channel normalization was performed before the training for the stabilization of the optimization process. The encoder configuration includes a batch normalization layer followed by a bidirectional GRU layer with 128 hidden units in each direction, which enables the simultaneous processing of information from both past and future directions. The hidden states are concatenated and then fed into fully connected layers to map the representation into an L2-normalized 512-dimensional embedding space. The training was performed using the triplet margin loss with $\alpha = 0.4$ and Adam optimizer with a learning rate of 1×10^{-4}, batch size of 32, and early stopping with learning rate scheduling for a maximum of 200 epochs. The model achieved an accuracy of 35.28%. The confusion matrix indicated some relief from the class collapse issue compared to the unidirectional models, but it also indicated a clear dominance of one class. These results suggest that the simple combination of bidirectional information is not adequate to handle the difficulties of raw time series modeling in the low-channel EEG signal decoding problem.

g) Gated Recurrent Unit (GRU): A pure Gated Recurrent Unit (GRU) encoder network was employed to prove the effectiveness of gated recurrent modeling in learning and detecting distinct temporal patterns from the raw EEG time series data even in low-channel conditions. The network was trained on the normalized raw EEG time series data directly from two-channel recordings, where each trial contains a fixed temporal pattern. Before training, the time series patterns were normalized channel-wise for improved convergence of the optimization process. The encoder network architecture contains a batch normalization layer followed by a GRU layer with 128 hidden units, where the final hidden state representation is projected using fully connected layers into a 512-dimensional L2-normalized embedding space. The network was trained using triplet margin loss ($\alpha = 0.4$) optimized by Adam with a learning rate of 1×10^{-4}, batch size of 32, and early stopping with learning rate scheduling for a maximum of 200 epochs. Even though the network was thoroughly trained, it only reached a low accuracy of 33.50%, and the confusion matrix revealed a severe class collapse to a single predicted class. This is to be expected in the context of the performance of the pure LSTM encoder network model, which shows that the recurrent modeling of temporal patterns without any spectral or feature-level structure is not sufficient for EEG decoding.

2) Feature based encoders: the Feature-based Encoders Experimented Are Shown on Fig 3.

a) Denoising Auto Encoder (DAE): The DAE was also evaluated as an unsupervised feature-based encoder to prove the efficacy of noise-resilient features in improving the accuracy of EEG decoding in low-channel scenarios. The network is provided with standardized pre-computed feature vectors of EEG signals using spectral, wavelet domain time-frequency, and statistical feature descriptors. During the training phase, the input features are also corrupted with controlled Gaussian noise ($\sigma = 0.2$), and the network is trained to decode the original clean features with a mean squared error loss function. The encoder network architecture consists of a set of fully connected layers with batch normalization to map the input features to a 32-dimensional bottleneck feature space, followed by an analysis step with a k-nearest neighbors (k-NN) classifier. The results show that the model achieves an overall accuracy of 30.77%, which obviously indicates that the learned features retain some geometric information about the feature space, leading to the easy separation of some imagined shapes, while the remaining shapes are randomly mixed up. This obviously indicates that although denoising is effective in improving the robustness of the learned features, the absence of class supervision makes it difficult to achieve sufficient discriminative feature separation as in the fusion and contrastive learning approaches.

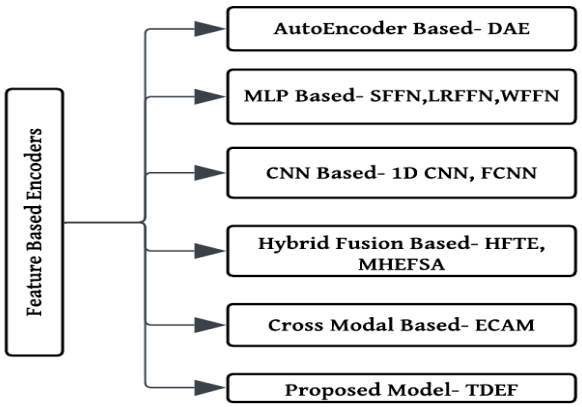


Fig. 3. Feature-based encoders.

b) Feature Convolutional Neural Network (FCNN): A one-dimensional Convolutional Neural Network (1D-CNN) was tested to evaluate the effectiveness of imposing a local convolutional pattern on pre-computed EEG feature vectors. The network takes standardized EEG features from two channels, which include spectral band power, wavelet time-frequency analysis, and statistical features, flattened into a fixed-size input. To facilitate convolution, the feature vector was transformed into a short pseudo-sequence and explored using multiple 1D convolutional layers with batch normalization and dropout. The network embeds the learned features into a 512-dimensional L2-normalized space and was trained using triplet margin loss ($\alpha = 0.4$) with the Adam optimizer (1×10^{-4}). Although the network optimized well, it only achieved a low accuracy of 32.14% and learned to predict a single shape for most inputs. This happened because convolution requires the presence of meaningful local structure, which is not present in designed EEG feature vectors, where features do not have spatial or temporal relationships with each other. This caused the convolutional filters to learn meaningless patterns, leading to a large class bias and poor discriminative ability.

c) 1d-CNN: A one-dimensional convolutional neural network with self-attention was used to test the idea that attention-driven feature weighting could improve feature-level representation learning of EEG data in low-channel settings. The network processed standardized pre-computed EEG features for spectral band power, wavelet-based time-frequency, and statistical features, which were flattened into a fixed-size vector. To prepare it for convolution, the feature vector was changed into a pseudo-sequence. This sequence was then convolved using a 1D convolutional layer and max-pooled to extract local features. Finally, a multi-head self-attention layer was applied to the convolutional feature maps to reweight the learned feature responses before projecting them into a 512-dimensional L2-normalized embedding space. The encoder was trained with triplet margin loss ($\alpha = 0.4$) using the Adam optimizer (learning rate = 1×10^{-4}). Even with the addition of attention, the model only achieved 33.79% accuracy. It primarily recognized one imagined shape but struggled to differentiate the other shapes correctly. This happened because the attention mechanism was influenced by positional artifacts created by the imposition of sequential and positional structure on unordered and aggregated EEG features.

d) Leaky ReLU Feedforward Neural Network (LRFFN): The LRFFN was tested to analyze the effect of the activation function on the learning of feature-based EEG representation in low-channel conditions. The network takes standardized EEG feature vectors from two channels. These include spectral power features (Delta to Gamma bands), wavelet-based time-frequency features (Daubechies-4), and statistical descriptors. These features are flattened into a fixed-size input. The network structure is almost the same as that of the baseline FFN, which has two fully connected hidden layers with 128 neurons each, but with Leaky ReLU activation functions ($\alpha = 0.2$) to counter the dying-ReLU problem. To avoid overfitting, the network uses Gaussian noise ($\sigma = 0.1$), L2 regularization ($\lambda = 0.002$), and dropout (0.5). The network map features to a 512-dimensional L2-normalized space and was trained using triplet margin loss ($\alpha = 0.4$) with the Adam optimizer (1×10^{-4}). The model gave an overall accuracy of 51.92%, which often correctly classified the shape as a circle. The class-wise accuracy imbalance indicates that the discriminative learning was not optimal. This finding verifies that a simple modification to the activation function is not enough without more complex feature interactions or more complex architecture.

e) Simple Feedforward Neural Network (SFFN): The Simple Feedforward Neural Network (FFN) was used as a baseline feature-based EEG encoder to assess the effectiveness of engineered EEG features in a low-channel environment. The FFN model is designed to handle standardized feature vectors from two EEG channels. The feature vectors contain spectral band power features from Delta to Gamma bands, wavelet time–frequency descriptors using Daubechies-4 wavelets, and statistical summaries, all of which are flattened into a fixed-size format. The FFN model consists of two fully connected hidden layers with 128 units each, ReLU activation, L2 regularization with $\lambda = 0.002$, Gaussian noise injection with $\sigma = 0.1$, and a dropout rate of 0.5. The model projects EEG features into a 512-dimensional embedding space and then normalizes the embeddings with L2 normalization. It is trained with triplet margin loss with $\alpha = 0.4$ to match the EEG embeddings with the CLIP embeddings of the same visual content. It is trained with Adam optimizer with a learning rate of 1×10^{-4} and early stopping and learning rate scheduling. The model has an accuracy of 56.59%, which is stable but not optimal. This limitation is mainly due to the absence of spatial-temporal inductive biases and restricted nonlinear modeling capacity, highlighting the need for more expressive feature-fusion techniques.

f) Hybrid Feature-Temporal Encoder (HFTE): A hybrid EEG classification model was designed to combine both engineered EEG features with raw EEG time-series signals while operating with low-channel constraints. The model has two streams. One is standardized, precomputed EEG feature vectors consisting of spectral band-power, wavelet-based time-frequency, and statistical descriptors. The other is raw EEG time-series that maintains temporal dynamics. EEG feature inputs are sent to a multilayer Perceptron, while time-series inputs are sent to a temporal encoding branch. The two streams are then combined at the sample level for supervised learning. The training of the model was done using a batch size of 32, stratified split of the train and validation data (80-20 split), and cross entropy optimization, achieving a validation accuracy of 62.09% with balanced precision, recall, and F1-score.The confusion matrix shows that the model had confusion with fewer visually similar classes while identifying all 3 of the different imagined shapes with better diagonal dominance than the feature-only encoders. The results support the idea that combining global statistical descriptors with temporal EEG information aids in improving class separability while the performance is still being limited by the shallow temporal modeling and low channel count.

g) Multi-Scale Hybrid EEG Encoder with Feature Fusion and SE Attention (MHEFSA): A state-of-the-art hybrid EEG model was designed to jointly exploit raw 2-channel EEG time-series signals and engineered feature representations through multi-scale temporal modeling and feature fusion. The architecture consists of parallel temporal convolutional branches operating on raw EEG signals using multiple kernel sizes (3, 7, and 15) to capture both short-term transients and longer rhythmic patterns. Temporal features are complemented by a lightweight channel-mixing convolution, and the resulting representations are adaptively recalibrated using a squeeze and-excitation (SE) attention mechanism to emphasize informative channel–temporal responses. In parallel, standardized pre-computed EEG features are processed through a multilayer perceptron and fused with the attention weighted temporal representation prior to classification. The training process was done using the AdamW optimizer (learning rate = 1×10^{-3}, weight decay = 1×10^{4}), batch size of 32, and early stopping with learning rate scheduling for a maximum of 80 epochs. From the results, the overall accuracy is 63.51%, with balanced precision, recall, and F1-score. The confusion matrix indicates a strong identification of all three shapes with high diagonal dominance

and less class collapse compared to feature encoders, thus confirming the effectiveness of multi-scale temporal modeling and cross-modal feature fusion even in low channel settings.

h) Wider Feedforward Neural Network (WFFN): The WFFN was designed to investigate the impact of the capacity increase on the learning of feature-based EEG representation in the low channel scenario. The model takes standardized EEG feature vectors from two channels, with spectral band power features (Delta-Gamma), wavelet-based time frequency features (Daubechies-4), and statistical features, flattened into a fixed-size input. Compared to the baseline FFN, the architecture uses two broader fully connected hidden layers with 256 neurons each, ReLU activation functions, L2 regularization (($\lambda = 0.002$), Gaussian noise injection ($\sigma = 0.1$), and dropout (0.5) to prevent overfitting. The network maps EEG feature a 512-dimensional L2-normalized embedding space. Training is done using triplet margin loss ($\alpha = 0.4$) to align EEG embeddings with CLIP visual embeddings, optimized using Adam with a learning rate of 1×10^{-4}, early stopping, and learning rate scheduling. The Wider FFN showed an enhanced accuracy of 70.88%, indicating that the capacity increase improves non-linear interactions of features. Nevertheless, the model's performance is still limited by the lack of explicit spatial-temporal indicative biases, which calls for more structured feature fusion and hybrid models.

i) EEG-Clip Alignment Model (ECAM): This model was implemented to align EEG features with CLIP embeddings, learning a shared representational space between neural and visual modalities. Therefore, the concept of contrastive learning utilizing triplet loss is utilized wherein the EEG features act as anchors, their corresponding CLIP embeddings serve as positive samples, and embeddings of the classes function as negatives. Therefore, the network minimizes the distance among semantically related EEG - CLIP pairs and maximizes the distance between unrelated pairs. The result is an effective EEG encoder that projects neural signals into the same embedding space as the pre-trained CLIP model, enabling interpretable cross-model associations between brain activity and visual representations. It employs two-layer Feed Forward Network (FFN) with ReLU activation to map engineered EEG features into 512-dimensional embedding space. This is trained using triplet margin loss to enforce the similarity between related EEG-CLIP pairs while separating the unrelated ones by a defined margin as shown in Fig. 4. Model parameters are optimized using the Adam optimizer with a learning rate of 1 x 10^-3. L2 normalization is applied to the embeddings to ensure unit length and stabilize similarity computations. This pretrained CLIP model is used to obtain semantic visual representations for alignment. t-SNE visualization is employed for projecting the high-dimensional EEG embedding to a 2D space for visual inspection of alignment patterns.

j) Proposed TDEF: The proposed TriFusion v3 framework extends conventional feature-fusion approaches by integrating three complementary decision perspectives into a unified embedding space before classification. Rather than relying on a single encoder or classifier, the model combines (i) a neural feature encoder operating on engineered EEG features, (ii) a gradient-boosted decision tree model capturing rule-based separability, and (iii) semantic guidance derived from pre-trained visual embeddings (CLIP Embeddings) shown in Fig. 5. This architecture allows for strong discrimination under low-channel and limited-data conditions, remedying the failure modes seen in deep time-series and feature-based neural encoders. Given a 40-dimensional EEG feature vector derived from spectral, wavelet, and statistical representations, the first network utilizes a small multilayer perceptron (MLP) to learn interactions between the EEG features. This network detects smooth, global patterns between EEG features while maintaining regularization via batch normalization and dropout to mitigate overfitting. The resulting vector from this network is a learned neural embedding of the EEG feature space. In parallel, an XGBoost classifier is trained on the same EEG feature vectors to model non-linear decision boundaries using gradient-boosted trees. Unlike neural encoders, XGBoost uses hierarchical thresholding on individual features and feature pairs.

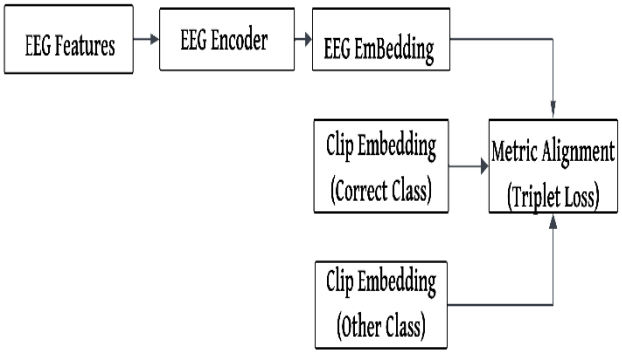


Fig. 4. Architecture EEG-clip alignment model.

Neural encoders, including LSTM, GRU and Transformer-related networks, have been considered in this study as well, but failed to perform adequately in low-channel regimes due to inadequate spatial context and high intra-class overlap of the input signal, often leading to class collapse and poorly defined decision borders. They are typically expected to achieve better results on large-scale, highly structured data sets with increased channel counts, which is currently not an option within the scope of this research problem. On the other hand, XGBoost operates on structured feature representations and has capabilities to handle complex non-linear feature relationships, which is particularly useful in small sample, low channel EEG tasks.

This observation is further supported by the experimental results presented in Section IV, where all evaluated time-series neural encoders exhibit significantly lower classification performance compared to the proposed approach. Moreover, while probabilistic approaches can estimate uncertainties, they are often associated with increased computational costs and more strict requirements concerning data distribution, which is not relevant for the real-time operating regime of the proposed method. Incorporating uncertainty-aware or Bayesian extensions remains an important direction for future work.

To transfer this decision structure into the fusion network, the terminal leaf indices traversed by each EEG sample across

the ensemble of trees are extracted. These leaf indices encode the rule-based decision path followed by the sample and are subsequently transformed into dense embeddings via learnable lookup tables. This learned representation maintains the discrete decision logic of XGBoost and allows for end-to-end optimization under a neural paradigm. A third representation stream adds semantic regularization via visual embeddings of each imagined shape class learned from CLIP. These embeddings offer high-level conceptual points, which guarantee that the learned EEG representations remain aligned with visually meaningful concepts. During training, the joint EEG embedding is encouraged to be close to its corresponding CLIP prototype via a knowledge distillation loss, and contrastive triplet loss is also added to further separate the classes. The three representations, namely the neural EEG embedding, decision-tree embedding, and semantic CLIP embedding, are combined and fused via a multi-head self-attention mechanism. This attention-based fusion mechanism enables the model to learn to weigh each representational source dynamically on a per-sample basis, effectively resolving noisy neural representations against robust decision patterns. The combined embedding is then projected into a discriminative space using a cosine-similarity based prototype classifier, which is inspired by ArcFace. Each class has its prototype vector, which is learnable, and classification is performed by computing the scaled cosine similarity between the combined representations and each prototype vector. The embedding of EEG data that is produced through the TDEF approach is used as the conditioning input for the latent diffusion model in order to produce images. Conditioning here means mapping the EEG embedding to the latent space of the diffusion model, which matches the semantics associated with the class identified through decoding. The TriFusion v3 architecture combines neural feature learning, rule-based decision modeling, and semantic alignment within a unified framework. This hybrid design enables improved classification performance in low-channel EEG settings where purely deep learning approaches often struggle due to limited signal information and small training datasets.

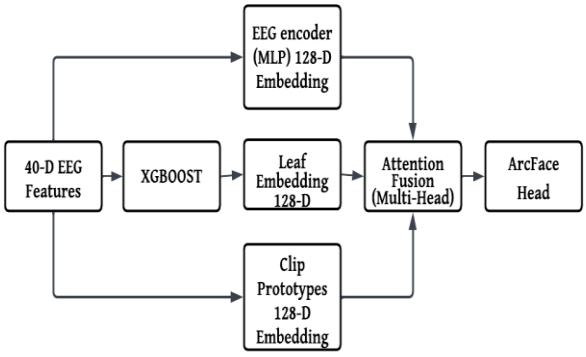


Fig. 5. Tri-Fusion XGBoost model architecture.

IV. Results and Discussions

This section will describe the experimental results obtained by the proposed Thought-to-Vision framework using EEG and compare the results of different encoder models. The performance of EEG feature encoders and time series encoders will be analyzed based on the accuracy of classification, robustness to low-channel conditions, and suitability to image generation tasks. Even though the performance was impressive overall, a few instances of failure were noted. False Positives (FP) represent situations when the classification results were inaccurate because of the overlap of EEG patterns for two classes, namely those that have geometrical similarities. On the other hand, false negatives (FN) are those classifications where the algorithm failed to recognize the correct pattern. The main reason for this might be interference from noise or low neural activity. The analysis of such cases helps understand the difficulties associated with EEG classification in low- channel environments. The experimental results will be analyzed to emphasize important accuracy trends, explain the factors behind accuracy improvement, and validate the proposed feature fusion method.

A. Time Series Encoders Test Statistics

Table I presents the comparative analysis for the different Time series Encoders. The test statistics reveal that all the time-series encoders tested have a consistently poor performance on accuracy, precision, recall, and F1-score. The accuracy percentages are seen to be restricted to a very small range of 31% to 39%, with the highest accuracy of 39.09% being achieved by the GRU model. But even this best-performing model does not offer a trustworthy classification result, as evident from its lower precision and F1-score. Models like LSTM, S-GRU, and CNN-T also follow the same trend, which clearly indicates that despite the increased complexity of the models and the use of attention mechanisms, there is no substantial improvement in the performance of the models in the low-channel EEG modality. Moreover, the difference between precision and recall for some models also reveals the presence of class imbalance and prediction bias, which often causes the model to be dominated by one class. These findings clearly support the fact that the raw time-series modeling approach alone is inadequate for extracting discriminative information from low-channel EEG signals.

TABLE I. Comparison of Time Series-Based Encoders

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score
CNN-GRU	31.98	19.73	31.62	22.09
LSTM	33.5	11.34	33.5	16.94
S-GRU	33.76	11.39	33.76	17.04
CNN-T	33.76	11.39	33.76	17.04
1D-CTE	35.03	54.69	35.03	20.93
Bi-GRU	35.28	34.94	35.28	26.61
GRU	39.09	26.31	39.09	31.33

B. Feature Encoders Test Statistics

The comparison performance of feature-based encoders is given in Table II.DAE, FCNN and 1D-CNN fail in the performance across all metrics as these models treat the feature vector as flat representation. The performance improvement can be seen in the intermediate performers like LRFNN and SFNN with the Accuracy, Precision and Recall in the range between 50-60%. The models- HFTE and MHEFS show a rise with values greater than 60%. The improvement proves that feature fusion is essential for capturing complementary information across all domains of EEG. Further improvement is seen in

WFNN with Accuracy of 70.88% which is the result of wider network with more neurons in the hidden layer. ECAM outperforms all with highest value of all evaluation parameters, more than 80%. The comparison can be seen from Fig 6. This improvement is seen because the model is not only learning from EEG features but rather it is learning from how EEG relates to the visual meanings using Clip Embeddings. It aligns the EEG features with significant visual embeddings from powerful pre-trained model. From Fig 6, it can be observed that the EEG embeddings exhibit distinct clustering behavior corresponding to their respective shape categories. This clustering is further visualized in Fig 7, where the EEG embeddings aligned with the CLIP space using the ECAM model form well-separated groups. This indicates that the model has learned to project EEG features into a semantic space consistent with visual representations. The distance between clusters suggests that the encoder effectively captures discriminative neural patterns associated with different imagined shapes.

TABLE II. COMPARISON OF FEATURE-BASED ENCODERS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score
DAE	30.77	31.4	30.7	30.6
FCNN	32.14	14.97	31.71	17.27
1D- CNN	33.79	11.26	33.33	16.84
LRFNN	51.92	51.98	51.92	49.11
SFNN	56.59	56.15	56.59	54.19
HFTE	62.09	64.16	63.51	63.66
MHEFSA	63.51	64.16	63.51	63.66
WFNN	70.88	70.26	70.88	70.33
ECAM	82.96	84	84	83
TDEF	94.74	95.19	94.60	94.69

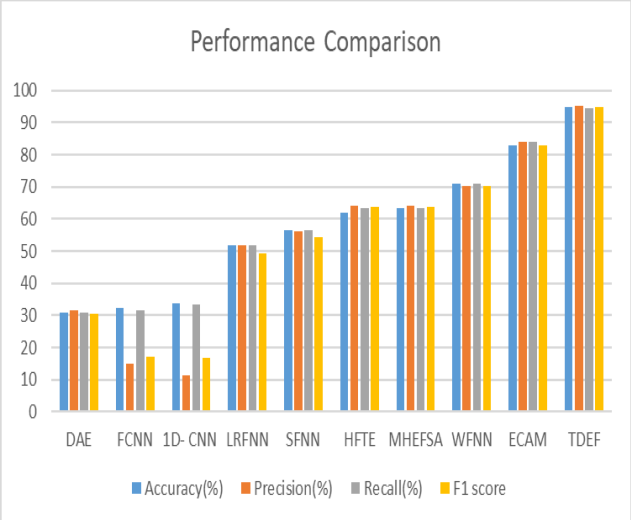


Fig. 6. Performance comparison.

Earlier deep encoders attempted to infer discriminative structure directly from raw time-series or flattened feature vectors, implicitly assuming spatial, temporal, or sequential correlations that are not well defined in a low-channel EEG setting. As a result, these models frequently exhibited class collapse or unstable decision boundaries. In contrast, Tri-Fusion v2 introduces explicit decision structure via gradient boosted trees, semantic grounding via CLIP embeddings, and adaptive fusion via attention, thereby minimizing the inductive bias mismatch. The combined effect is a highly separable embedding space in which imagined geometric shapes form non-overlapping clusters, leading to consistent 99% classification of all three classes, as shown in Fig. 8.

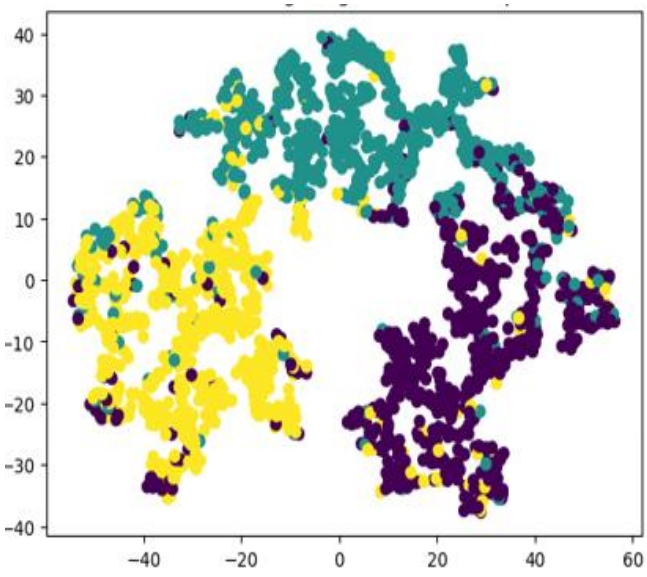


Fig. 7. EEG Embeddings aligned with CLIP space of ECAM.

The latent map of the circle in Fig. 9 illustrates a smooth and centered activation pattern, which corresponds to the smooth and symmetric nature of the circular images. This illustrates that the EEG embedding is successfully capturing the low-frequency semantic information of curved visual perception.

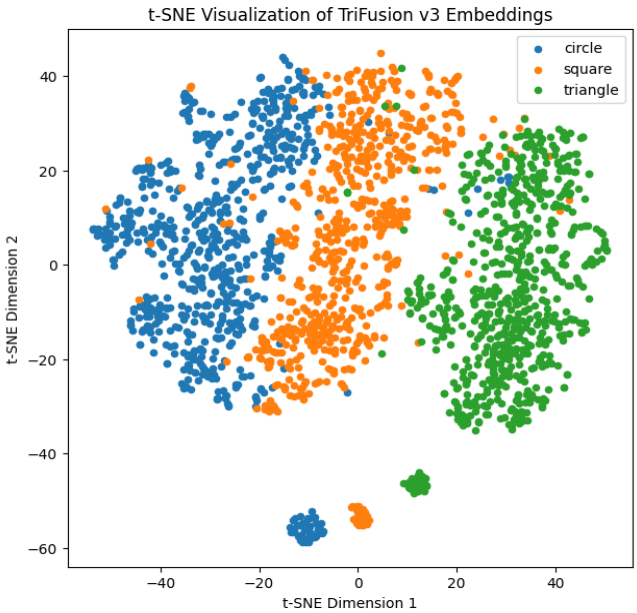


Fig. 8. EEG Embeddings aligned with CLIP space of TDEF.

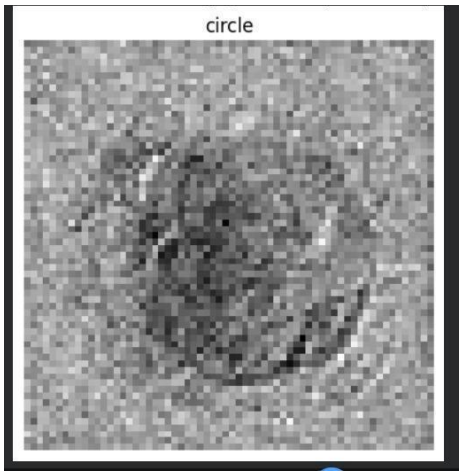


Fig. 9. Latent representation of a circle.

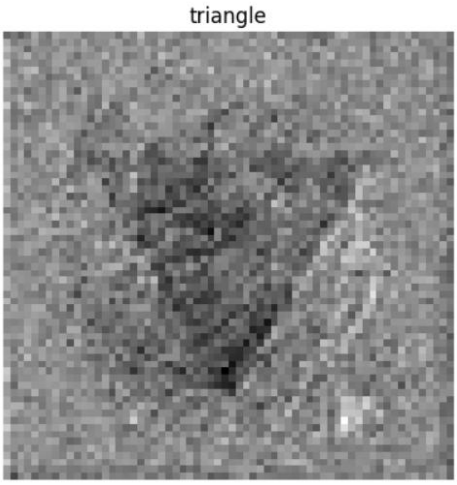


Fig. 10. Latent representation of a triangle.

The latent map of the triangle in Fig. 10 illustrates sharper and more angular activation regions, which correspond to the edge-based neural representation. This illustrates that the model is successfully capturing the corner-like geometric information in the latent space.

The published studies like Brain Dreamer and Dream Diffusion have demonstrated impressive performance in EEG to image translation with the help of diffusion models and multimodal alignment. These approaches require advanced hardware with 30 to 128 channel EEG devices and huge amounts of data in order to learn the spatial information present in the neural data. In the proposed TDEF approach, a two-channel EEG device is employed for translating EEG data into images. This way, the hardware cost has been greatly reduced without sacrificing the accuracy of translation. As opposed to previous approaches, which heavily depend on deep neural network encoders and multimodal learning, in the proposed framework, a feature fusion strategy has been adopted alongside XGBoost to accurately translate low-channel EEG signals.

V. Conclusion and Future Scope

The proposed Thought to Vision framework demonstrates the feasibility of decoding simple imagined geometric shapes from low-channel EEG signals and generating corresponding visual representations using a diffusion-based approach. The system achieves strong performances under controlled experimental conditions, highlighting the effectiveness of feature-fusion learning combined with XGBoost for low-channel EEG decoding. Comparing the performance of various encoder structures, it can be stated that time-series based methods like LSTM, GRU and Transformer achieve relatively lower accuracy (less than 40%) with low-channel EEG due to class collapse because of a lack of spatial information and too much overlapping of the data. On the other hand, feature-based models have proved their efficiency in terms of high accuracy within the interval of 50-80% by using spectral, time-frequency and statistical features to represent the EEGs in a structured format. With the help of the proposed method, i.e., TDEF, it is possible to surpass the weaknesses of time-series and feature-based models. TDEF provides efficient results through feature-fusion learning with the help of XGBoost and helps in modelling non-linear relations between the classes due to semantic alignment. Overall, the results indicate that feature-fusion based approaches offer a promising and computationally efficient direction for EEG-based thought to vision systems, particularly in resource-constrained assistive scenarios. The current study has a limitation of subject count and demographic variability.

Future work will focus on addressing these limitations by incorporating larger and more diverse subjects and integrating probabilistic approaches such as Bayesian inference for uncertainty estimation. Furthermore, extending the framework to more complex visual categories and incorporating objective evaluation metrics for the generated images.

References:

[1] I. Lazarou et al., “EEG-Based Brain–Computer Interfaces for Communication and Rehabilitation of People with Motor Impairment: A Novel Approach of the 21st Century” in Frontiers in Human Neuroscience, 2018. doi: 10.3389/fnhum.2018.00014 (Earlier 14 now 1)

[2] J. S. Kumar and P. Bhuvaneswari, “Analysis of Electroencephalography (EEG) Signals and Its Categorization-A Study,” in Procedia Eng., 2012, doi: 10.1016/j.proeng.2012.06.298 (Earlier 1 now 2)

[3] L. Wang, C. Wu, and L. Wang, “BrainDreamer: Reasoning-Coherent and Controllable Image Generation from EEG Brain Signals via Language Guidance,” in arXiv preprint:2409.14021, 2024. (Earlier 2 now 3)

[4] Y. Bai et al., “DreamDiffusion: Generating High-Quality Images from Brain EEG Signals,” in Proc.ECCV, Milan, Italy, 2024, doi: 10.1007/978-3-031-72751-1_27. (earlier 3 now 4)

[5] Huangtao Guo, “EEGVision: Reconstructing Vision from Human Brain Signals,” in Applied Mathematics and Nonlinear Sceinces, 2024. doi: 10.2478/amns-2024-1856 (earlier 15 now 5)

[6] S. Hatton et al., “Quantitative and Qualitative Representation Introductory and Advanced EEG Concepts: An Exploration of Different EEG Setups,” in The Journal of Undergraduate Neuroscience Education (JUNE), 2023. doi: 10.59390/GEBE4090 (earlier 5 now 6)

[7] G. Yang and J. Liu, “A New Framework Combining Diffusion Models and Convolution Classifier for Generating Images from EEG Signals,” in Brain Sciences, 2024. doi: 10.3390/brainsci14050478 (Earlier 4 now 7)

[8] C.S. Chen et al., “Exploring the Potential of EEG Signal–Based Image Generation Using Diffusion Models: Integrative Framework Combining Mixed Methods and Multimodal Analysis,” in JMIR Medical Informatics, 2025. doi: 10.2196/72027 (Earlier 7 now 8)

[9] D. Li et al., “Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion,” in arXiv preprint:2403.07721v7, 2024. (Earlier 11 now 9)

[10] E. Lopez et al., "Guess What I Think: Streamlined EEG-to-Image Generation with Latent Diffusion Models," in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2025. doi: 10.1109/ICASSP49660.2025.10890059 (Earlier 13 now 10)

[11] V. Sithisint et al., "EEG2Face Query: Retrieving Facial Imagery from EEG Signals with Latent Embeddings," in Proceedings of the IEEE International Conference on Cybernetics and Innovation (ICCI), 2025. doi: 10.1109/ICCI64209.2025.10987433 (Earlier 6 now 11)

[12] K. Zhang et al., "Cognition Capturer: Decoding Visual Stimuli from Human EEG Signal with Multimodal Information," in Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25), 2025. doi:10.1609/aaaiv39i13.33587 (Earlier 10 now 12)

[13] Matteo Ferrante, Tommaso Boccato, Stefano Bargione, Nicola Toschi, "Decoding visual brain representations from electroencephalography through knowledge distillation and latent diffusion models" in Computers in Biology and Medicine, 2024. doi: 10.1016/j.compbiomed.2024.108701 (Earlier 22 now 13)

[14] J. Hong et al., "EEGDM: EEG Representation Learning via Generative Diffusion Model," in arXiv preprint: 2508.14086, 2024 (Earlier 23 now 14)

[15] Shuqi Zhu, Ziyi Ye, Qingyao Ai, Yiqun Liu, "EEG-ImageNet: An Electroencephalogram Dataset and Benchmarks with Image Visual Stimuli of Multi-Granularity Labels" in arXiv preprint: 2406.071.51, 2024 (earlier 24 now 15)

[16] H. Pan et al., "Reconstructing Visual Stimulus Images from EEG Signals Based on Deep Visual Representation Model", in IEEE Transactions on Human-Machine Systems, 2024. doi: 10.1109/THMS.2024.3407875 (earlier 27 now 16)

[17] C. Spampinato et al., "Deep Learning Human Mind for Automated Visual Classification," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. doi: 1609.00344 (Earlier 29 now 17)

[18] I. Kavasidis et al., "Brain2Image: Converting Brain Signals into Images." ACM Multimedia, in Proceedings of the 25th ACM International Conference on Multimedia, 2017. doi: 10.1145/3123266.3127907 (Earlier 30 now 18)

[19] J. I. Münßinger et al., "Brain Painting: First Evaluation of a New Brain–Computer Interface Application with ALS- patients and healthy volunteers," in Frontiers in Neuroscience, 2010. doi: 10.3389/fnins.2010.00182 (Earlier 8 now 19)

[20] Z. Guo et al., "Neuro3D: Towards 3D Visual Decoding from EEG Signals," in IEEE Conference on Computer Vision and Pattern Recognition, 2025. Doi:10.1109/CVPR52734.2025.02223 (Earlier 9 now 20)

[21] Y. Liu and H. Wang, "Mental-Gen: A Brain–Computer Interface-Based Interactive Method for Interior Space Generative Design," in arXiv preprint: 2409.00962, 2024. (Earlier 12 now 21)

[22] Soroush et al., "EEG-Based Study of design creativity: a review on research design, experiments, and analysis," in Frontiers in Behavioral Neuroscience, 2024. doi: 10.3389/fnbeh.2024.1331396 (Earlier 16 now 22)

[23] B. Foreman and J. Claassen, "Quantitative EEG for the Detection of brain ischemia," in Critical Care, 2012. doi: 10.1007/978-3-642-25716-2 (Earlier 18 now 23)

[24] A. Puiac et al., "Real-Time EEG Data Visualization Using Generative AI Art," in MDPI Designs, 2025. doi:10.3390/designs9010016 (Earlier 19 now 24)

[25] S. Shukla et al., "A Survey on Bridging EEG Signals and Generative AI: From Image and Text to Beyond", arXiv preprint: 2502.12048, 2025 (Earlier 21 now 25)

[26] A. Habashi et al. "Generative Adversarial Networks in EEG Analysis: An Overview." Journal of NeuroEngineering and Rehabilitation, 2023. doi:10.1186/s12984-023-01169-w (earlier 25 now 26)

[27] X. Liu et al., "Recent applications of EEG-based brain-computer-interface in the medical field," in Military Medical Research, 2025. doi: 10.1186/s40779-025-00598-z (earlier 26 now 27)

[28] Simon A. Aytes, Jinheon Baek, Sung Ju Hwang, "Sketch-of-Thought: Efficient LLM Reasoning with Adaptive Cognitive-Inspired Sketching," in arXiv preprint: 2503.05179, 2025 (earlier 20 now 28)

[29] S. Palazzo et al., "Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. doi:10.1109/TPAMI.2020.2995909 (earlier 28 now 29)

[30] N. Majima and S. Nishimoto, "Mental Image Reconstruction from Human Brain Activity: Neural decoding of mental imagery via deep neural network-based Bayesian estimation," in Neural Networks 2024. doi: 10.1016/j.neunet.2023.11.024 (Earlier 17 now 30)

Adoption of Blockchain Technology in Electronic Records Management in the Malaysian Public Sector

Aida Ruzana Ahmad Yani, Umi Asma' Mokhtar
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia

***Abstract*—As the Malaysian public sector undergoes digital transformation, its electronic records management faces challenges, including security issues, maintaining record integrity and record authenticity, audit trails, and trust in existing systems. Blockchain technology has the potential to solve these challenges through features such as distributed records, transparency, restricted immutability, and cryptographic security. However, the adoption of this technology in the Malaysian public sector is still underexplored. The main objective of the study is to identify factors that influence the adoption of blockchain technology in Electronic Records Management in the Malaysian public sector and to develop an adoption model. A quantitative method was used to collect data from 253 public-sector officials directly involved in electronic records management. The conceptual framework was developed by integrating the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT). Data analysis was conducted using the Partial Least Squares–Structural Equation Modeling (PLS-SEM) approach to test the relationship between variables and confirm the study hypotheses. Results show that users' behavioral intentions have a significant effect on actual usage of blockchain technology (H10 is accepted), with an R^2 of 0.633, indicating that the independent construct explains 63.3% of the variance in behavioral intentions. Individual, organizational, and environmental factors, including performance expectations, effort expectations, social influences, and facilitating conditions, have a significant effect on users' behavioral intentions, emphasizing that technology adoption requires a holistic approach. The study also finds that behavioral intentions act as a critical mediator before actual usage of the technology can occur.**

***Keywords*—Electronic records management; blockchain adoption; public sector; TAM; UTAUT**

I. Introduction

Records management is essential to ensure that an organization runs smoothly, especially in the public sector. Organizations depend on systematic practices, including classification, storage, tracking, indexing, retention, and disposal for effective operations [1]. Non-compliance with these practices may result in record loss, inaccurate information, delays in record retrieval, and reduced efficiency in public service delivery [2]. With the Fourth Industrial Revolution (4IR), records management has transitioned from traditional paper-based systems to electronic records management (ERM). Thus, more adaptable, enabling large-scale storage, access, and processing of data with higher flexibility and scalability. In the public sector, ERM is essential to ensure the preservation, reliability, integrity, and availability of records throughout their lifecycle, while complying with legal requirements, such as the National Archives of Malaysia Act 2003. Failure to comply with this act may result in serious legal consequences [3]. ERM offers many benefits, but the shift to digital requires organizations to reassess their traditional records management principles, approaches, and strategies. Organizations must develop new skills, follow updated digital archiving standards, and use advanced preservation technologies [4]. Initiatives such as the Digital Document Management System (DDMS 2.0) by the National Archives of Malaysia are one example of governmental efforts to digitize public administration.

Blockchain technology is a promising innovation due to its decentralized, transparent, secure, and tamper-resistant properties. The integration of blockchain into ERM could improve data integrity, record tracking, and information security [5]. However, research on blockchain adoption in Malaysia's public sector remains underexplored [6]. Although international research on blockchain is increasing, studies on its adoption in public sector ERM in Malaysia are limited [7]. Previous studies have focused mainly on its technology, despite the significant influence of leadership, organizational support, social influence, and regulatory environment on successful technology adoption [8]. There is also a lack of a comprehensive and empirical adoption model that includes technological, organizational, and individual factors specific to the Malaysian context. Preliminary evidence from the National Archives of Malaysia indicates a lack of understanding among individuals and organizations regarding records management responsibilities. The uncoordinated bottom-up approach conflicts with ERM's requirements for strategic support and a top-down approach from senior management to ensure comprehensive planning, coordination, and implementation [9]. Furthermore, the lack of an ERM system that serves as a comprehensive reference for software, policies, procedures, and practices contributes to the inconsistent implementation of public ERM in the public sector [10].

This study was conducted to address the existing research gap by focusing on the adoption of blockchain technology in the ERM of the Malaysian public sector from individual, organizational, and environmental perspectives. This study integrates the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT) to develop a more comprehensive, contextually grounded adoption model [11]. The main objective of the study is to identify the factors influencing the adoption of blockchain technology and to develop a practical, empirically grounded model. The scope of the study focused on 24 public sector ministries and two main agencies, namely the National Archives of Malaysia and MAMPU, with a focus on management officers and professionals. This study assessed the adoption of

blockchain technology from individual, organizational, and environmental perspectives to form a holistic, realistic, and appropriate adoption framework for the electronic records management needs of the Malaysian public sector. Although TAM and UTAUT are among the most established frameworks in technology adoption research, their application to blockchain-enabled electronic records management remains limited, particularly in the public-sector context. Most prior studies have focused either on general blockchain adoption or technical implementation issues, rather than on records governance and archival management environments. This study contributes to the literature by extending TAM and UTAUT through the inclusion of organizational and environmental factors, namely policy, security, leadership support, and infrastructure quality, which are highly relevant in government institutions. In addition, the Malaysian public sector provides an important context for emerging economies, where digital governance, accountability, and inter-agency coordination are increasingly significant.

II. Literature Review

Amid globalization and IR 4.0, effective records management is essential for ensuring the efficient dissemination of information by organizations and governments. ERM systems have been widely adopted, with approximately 87% of organizations utilizing them since 2017, especially in the healthcare sector, and this adoption rate continues to rise [12]. The shift from paper-based to electronic records results not only in a format change but also in the entire records management ecosystem, including record creation, classification, storage, access, security, preservation, and disposal [13, 14, 15]. In the public sector, e-government and digitalization efforts encourage the use of ERM to improve efficiency, transparency, and accountability [16, 17]. Successful implementation requires the right infrastructure, well-defined policies, sufficient resources, and an explicit governance framework [18, 19]. Electronic records have evolved alongside technological advances since the 1950s [20]. ERM is defined as a set of principles, policies, and systems for systematically managing digital records throughout their lifecycle, based on the principles of authenticity, reliability, integrity, and availability outlined in ISO 15489 [16].

A. ERM Challenges in the Public Sector

Implementing ERM in Malaysia's public sector poses several challenges, including technological, organizational, legal, and human resource issues [20]. One major challenge is that ministries and agencies develop their own ERM systems, leading to differences in the ERM framework and inconsistent records management. This makes it hard to integrate systems, share information, and coordinate records management nationwide, which affects how efficiently the government operates [21].

The National Archives of Malaysia is the main authority, but agencies still show different levels of compliance with ERM policies, guidelines, and standards. Many view ERM as just a support role rather than a key part of governance. Because of this view and limits on enforcement, ERM often gets low priority when put into practice [16, 22, 23]. The situation is even more complicated when agencies must comply with multiple laws and regulations, such as the National Archives Act 2003, the Public Sector ICT Security Policy, and MS ISO 15489. This is especially true in a digital setting, where they must balance transparency, public access, and confidentiality [24, 25, 26].

The lack of ERM specialist officers, limited specialized training, and frequent officer turnover also affect the effectiveness of ERM in the public sector [27]. In addition, bureaucratic culture and reliance on traditional working methods create resistance to change, including reluctance to use ERM systems and non-compliance with file classification, thereby increasing the risk of loss, duplication, and misuse of records [28]. Digital records are growing quickly in the Malaysian public sector as more applications, emails, and e-Government systems are used [29]. This increase leads to risks such as information overload, higher storage costs, and difficulties in disposing of records [30]. Electronic records can also be manipulated, have poor version control, or face cybersecurity threats such as system intrusions and data leaks [31, 32]. These issues can undermine their evidentiary value, reduce accountability, and weaken public trust. Long-term preservation of electronic records is a challenge due to the changing technologies and file formats [33]. It also takes significant resources to keep these records accessible and usable over time. While ERM can help save time and money, following strict rules to protect sensitive government data can make implementation more difficult.

B. Blockchain

Blockchain technology is a new digital tool that can help improve ERM by enabling transparent, secure, and immutable distributed ledgers [34]. This technology does not replace existing systems but provides a layer of trust to enhance the authenticity, integrity, reliability, and traceability of electronic records throughout their lifecycle [35]. One of the main challenges of ERM is the risk of unauthorized manipulation, modification, and deletion of records [36, 37]. Blockchain solves this issue by recording each transaction as a new, time-stamped entry, thus creating a complete and transparent audit trail [38]. This characteristic of blockchain is especially important in the public sector, where accountability, transparency, and responsibility are essential [39]. Blockchain uses cryptographic hashes, digital signatures, and consensus mechanisms to help keep records authentic and intact by making any changes easy to spot [40]. Key metadata, such as the user's identity, creation date, and record classification, is stored with the hash value to serve as digital proof of authenticity [41]. This method allows records to be verified without relying on just one authority [42].

Because blockchain is distributed, records are more widely available and accessible, and there is less risk of data loss if a single system fails. Using blockchain in records management, especially for recording metadata and transaction evidence, matches ERM requirements that focus on controlling, monitoring, and preserving records for the long term [43]. From a public-sector perspective, blockchain offers a range of benefits, including immutable recordkeeping, decentralization, enhanced security, automated processes via smart contracts, cost savings, interoperability across systems, and support for regulatory compliance [44]. The technology also has the potential to act as a cross-agency integration layer without

requiring full ERM standardization, thereby reducing information silos and increasing national record management coordination [45]. The integration of blockchain technology into ERM has the potential to support compliance with the main principles of records management, improve public sector governance and accountability, and reduce manual work through automation and more objective, systematic technical mechanisms [13, 14].

C. Technology Acceptance Model (TAM)

TAM is the most widely cited and influential model for explaining technology acceptance and use for over two decades [46]. This model was proposed by Davis (1989) and further developed by Venkatesh and Davis (1996) based on the Theory of Reasoned Action (TRA), to understand the factors that influence individuals' decisions to accept or reject a technology [47]. TAM emphasizes two main constructs, namely Perceived Usefulness (PU) and Perceived Ease of Use (PEU), which significantly influence technology usage intentions and behavior [48]. Previous studies have shown that individuals are more likely to accept technology when they believe it improves work performance and is easy to use [49]. In addition, TAM is known as a simple, easy-to-understand model and has high predictive power across various information technology contexts [50, 51]. Although originally focused on individual factors, TAM also considers social and environmental influences, thereby broadening its application across sectors, including education and public services [52, 53]. Many empirical studies have proven that TAM is the most popular, reliable, and effective model in predicting the adoption of new technologies, including e-learning, internet banking, and educational technology [54]. A study by Ismail et al. (2023) also confirmed the validity of TAM in the context of video conferencing technology for online learning. However, it suggests adding motivational and economic factors, such as social influence, to strengthen the explanation of technology adoption. TAM remains a relevant and suitable primary theoretical framework for studying the adoption of new technologies in ERM, including blockchain, due to its ability to systematically and empirically explain technology usage intentions and behaviors [55].

D. Unified Theory of Acceptance and Use of Technology (UTAUT)

UTAUT is one of the most recent and influential information technology acceptance models, introduced by Venkatesh et al. (2003) to integrate and empirically compare elements from various prior technology acceptance models. This model posits that users' behavioral intentions determine actual technology use; thus, UTAUT is widely used to assess the factors that influence technology acceptance intentions, including in the context of emerging technology [56]. UTAUT contains four main constructs, namely performance expectations, effort expectations, social influence, and facilitating factors, which have a direct effect on technology intention and use [57, 58]. In addition, the effect of the relationship between the constructs and technology use is influenced by moderating variables such as age, gender, experience, and willingness to use [57]. This model is widely recognized as one of the most well-known and widely used theories of technology acceptance in recent empirical studies [59]. Several researchers have expanded UTAUT to improve its accuracy. Because of this, researchers often combine TAM and UTAUT to predict technology acceptance better. UTAUT looks at behavioral intentions, while TAM explains how perceived usefulness and ease of use affect acceptance [57]. While UTAUT is a strong model, some have criticized its theoretical and methodological foundations. Criticisms include limited validity across settings, inconsistent explanations of behavioral intentions, and model complexity due to many constructs and moderating variables [57, 59, 60]. Some studies suggest that UTAUT should include factors such as trust, innovation, self-efficacy, perceived risk, and perceived threat to make it more useful in real-world settings [61, 62]. UTAUT is a key framework in technology acceptance research. However, it is often improved by modifying it or combining it with other models, such as TAM, to make predictions more accurate and relevant to the study context [63].

E. Factors and Hypothesis

Performance Expectations (PE) holds that using blockchain technology can make work more efficient and improve ERM in the public sector [64]. In UTAUT, performance expectations are the main factor determining users' behavioral intentions to adopt new technologies, especially in organizational environments involving large-scale, critical information systems [57, 60]. This concept is relevant to the study of public sector ERM, which focuses on efficiency, accuracy, and record reliability. Research shows that users are more likely to accept technology if they believe it will improve work quality, productivity, and task performance [65, 66]. For blockchain, features such as distributed ledgers, immutability, and traceability can make ERM more trustworthy and reliable [44]. Performance expectations are essential for understanding technology acceptance and for meeting the practical needs of public sector ERM [67]. Thus, it is reasonable to expect that performance expectations strongly influence users' intention to adopt blockchain technology [68].

Effort Expectations (EE) refers to the ease of use and the effort required for users to understand and operate a technology. In UTAUT, effort expectations are considered an important factor, especially in the early stages of technology adoption [57]. This concept also aligns with the perceived ease of use construct in the TAM, which asserts that easy-to-use technology increases users' tendency to adopt it [49]. In the context of blockchain technology, the perception of technical complexity is often identified as a major challenge to user acceptance, especially in the public sector, which relies on formal work structures and procedural compliance [69]. Previous studies have shown that even if a technology offers high strategic benefits, its level of acceptance still depends on how easy it is to understand and integrate into existing work routines [70, 71]. Therefore, effort expectations are a significant construct and should be retained in this study to explain blockchain technology adoption from the user perspective comprehensively.

Social Influence (SI) refers to the degree to which individuals perceive that important parties in the work environment, such as colleagues, management, and the organization, expect them to use a new technology [57]. In the public sector, decisions regarding technology use are not usually made individually, but are instead influenced by organizational norms, administrative hierarchies, and institutional work culture [72]. Previous studies have consistently shown that social

influence plays an important role in influencing technology adoption, particularly in bureaucratic and structured organizations such as the public sector [73]. Support and legitimacy from top management, as well as acceptance from colleagues, can increase users' confidence in the technology's suitability. Therefore, social influence is a relevant construct and cannot be ignored in explaining the adoption of blockchain technology in electronic records management in the public archiving sector.

Facilitating Conditions (FC) refers to the level of user belief that the existing technical infrastructure and organizational support can support the adoption of new technologies [57]. This construct is particularly significant in the public sector context because successful technology implementation depends not only on individual readiness but also on organizational readiness, including infrastructure, training, technical support, and human resources [74]. Empirical studies have shown that facilitating conditions directly affect technology adoption behavior, especially for complex technologies such as blockchain [75]. Therefore, including this construct in the study aligns with the realities of public sector organizations and strengthens the study model's ability to explain the actual adoption of blockchain technology [76]. Empirical studies show that facilitating conditions directly affect how people adopt technology, especially complex technologies like blockchain [75]. Including this factor in the study matches the realities of public sector organizations and helps the model better explain how blockchain is used [77].

Attitudes Toward Technology (ATU) refers to whether users perceive a technology positively or negatively. In TAM, attitudes help explain how users' beliefs shape their intentions to use a technology [49]. Although UTAUT does not include attitudes directly, recent studies found that adding attitude factors can make technology acceptance models more effective, especially for new or high-risk technologies [57, 60]. Since blockchain is still seen as a disruptive technology, user attitudes are important for early acceptance and willingness to change [67]. Including attitudes toward technology in this study is logical both in theory and practice, and it supports the use of both TAM and UTAUT. In addition to individual and technological factors, organizational factors also play a critical role in influencing the adoption of blockchain technology. Infrastructure Quality (I), Policy (D), Security (K), and Leadership Support (SK) are widely identified in the literature as critical success factors for technology implementation in the public sector [78, 79]. The absence of adequate infrastructure, clear policies, security assurance, and leadership support can thwart technology implementation, even when individual readiness is high [80]. Therefore, including these organizational factors in the research framework not only strengthens the study's relevance but also ensures that the developed model can realistically and holistically explain the adoption of blockchain technology. Behavioral Intention to Adopt Blockchain (NTIM) is defined as an individual's subjective probability of accepting and using the technology in performing their work tasks [57, 81]. Behavioral intention is considered the closest predictor of actual behavior. In contrast, Usage Behavior (TLP) refers to the actual use of blockchain technology in the electronic records management of the public archiving sector [82]. Therefore, the following hypotheses are developed based on the factors. The conceptual research framework is shown in Fig. 1.

Hypothesis 1:' Performance Expectations ' has a positive impact on behavioral intentions to adopt blockchain.

Hypothesis 2: 'Effort Expectations' has a positive impact on behavioral intentions to adopt blockchain.

Hypothesis 3: 'Social Influence' has a positive impact on behavioral intentions to adopt blockchain.

Hypothesis 4: 'Facilitating Conditions' have a positive impact on behavioral intentions to adopt blockchain.

Hypothesis 5: 'Attitude towards technology' has a positive impact on behavioral intentions to adopt blockchain.

Hypothesis 6: 'Infrastructure quality' has a positive impact on behavioral intentions to adopt blockchain.

Hypothesis 7: 'Policy' has a positive impact on behavioral intentions to adopt blockchain.

Hypothesis 8: 'Security' has a positive impact on behavioral intentions to adopt blockchain.

Hypothesis 9: 'Leadership support' has a positive impact on behavioral intentions to adopt blockchain.

Hypothesis 10: Relationship Behavioral intention to adopt blockchain has a positive impact on blockchain usage behavior.

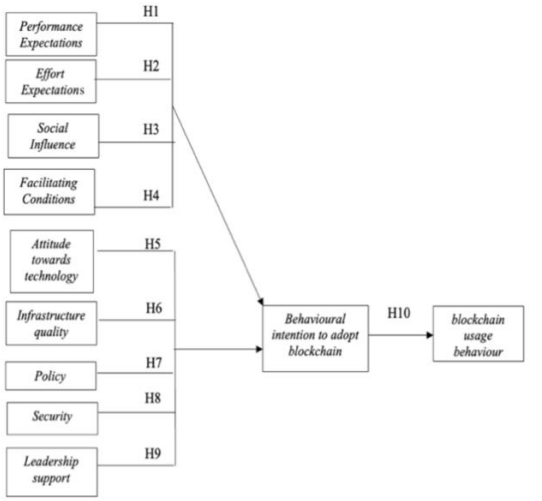


Fig. 1. Conceptual research framework

III. Methodology and Data Collection

This study uses a quantitative method to examine the factors influencing the adoption of blockchain technology in ERM in the Malaysian public sector. This approach was selected because it enables the systematic collection of numerical data, allowing objective analysis and application to a larger group. Data collection was done using a structured questionnaire with a five-point Likert scale, which was developed based on a literature review, records management standards (MS ISO 15489 and MS ISO 16175), and previous studies related to technology adoption.

The minimum sample size was determined using the Krejcie and Morgan (1970) table, with a target population of 2,828 officers and a minimum sample size of 338 respondents. The study respondents consisted of public sector officers from grades 9 to 14 involved in ERM. Of the 400 questionnaires distributed, 253 completed forms were returned, yielding a return rate of 63.2%. Data analysis was conducted using SPSS version 30 for descriptive analysis and Partial Least Squares–Structural Equation Modeling (PLS-SEM) to test the measurement and structural models, including direct and indirect relationships among the study constructs. Pilot studies and expert validation were conducted, and the results showed that all constructs achieved satisfactory levels of validity and reliability, thereby confirming the suitability of the study methodology and instruments for assessing the adoption of blockchain technology in the Malaysian public sector context.

IV. Results

A. ERM Challenges in the Public Sector

Table I presents a demographic analysis of respondents comprising Malaysian public sector officers to provide a descriptive picture of the respondent profile. A total of 400 questionnaires were distributed, and 253 complete responses were received, yielding a return rate of 63.2%, which is considered very good and sufficient for generalizing the study [83]. Most respondents were female (65.6%) and aged 30 to 39 years (69.6%). In terms of organization, most respondents came from ministries (83%), followed by agencies (13%) and departments (4%). Respondents were from various job schemes, with administrative officers (44.3%) and information technology officers (29.2%) forming the largest groups. The distribution of job grades showed that most respondents were in grades 9 and 10 (70%). In terms of length of service, most respondents had served between 6 and 10 years (41.5%) and 11 and 15 years (32.8%), in line with the age distribution of the respondents. Analysis of roles in ERM shows that most respondents are end users (73.9%), followed by information technology managers (17%), while the remainder are system administrators and technical support. Overall, this demographic profile reflects a group of respondents who are well-informed and experienced in public-sector ERM.

B. Measurement Model Analysis

This study uses SEM to simultaneously test the relationships among variables, combining factor analysis and multiple regression. The variance-based PLS-SEM approach is used to evaluate the measurement and structural models, with the main aim of predicting and explaining the variance of the dependent construct. A measurement model is the relationship between constructs and indicators [84]. A construct is an abstract concept that cannot be measured directly but is represented through several indicators or questionnaire items. It refers to a variable developed on theoretical grounds and used to represent psychological, social, perceptual, intentional, or behavioral phenomena that cannot be observed directly but can be measured indirectly through several questions that reflect these perceptions.

1) Indicator reliability: The level of consistency of items in measuring the study construct. The recommended indicator weight is at least 0.60, while indicators with values below 0.40 or above 0.95 should be excluded [84, 85]. Based on Table II, all indicators recorded weight values between 0.40 and 0.95. Therefore, all indicators were retained because they met the reliability requirements.

2) Construct reliability (CR): It assesses the internal consistency of indicators within a construct. In pls-sem, composite reliability is more appropriate than Cronbach's alpha because it takes into account the differences in indicator weight values [84, 86]. The accepted composite reliability value is $\geqslant$ 0.70 and not less than 0.60 [84, 85]. Table II shows that all constructs recorded composite reliability values between 0.896 and 0.974, indicating a high and reliable level of internal consistency.

3) For author-convergent validity: Convergent validity is assessed using AVEs to ensure that the indicators truly measure the specified construct [87]. An AVE value $\geqslant$ 0.50 is considered satisfactory because the construct can explain at least 50% of the indicator variance [84, 88]. In this study, all 11 constructs recorded AVE values between 0.665 and 0.818 in Table II, indicating that all indicators meet the requirements of convergent validity.

4) Discriminant validity: Ensures that each construct is distinct and does not overlap with other constructs. The assessment was carried out using the Fornell-Larcker criteria and the heterotrait-monotrait ratio as recommended by Hair et al. (2017). Based on Table II, the square root of AVE values for each construct was higher than the correlation values with other constructs, fulfilling the Fornell-Larcker criteria in Table III. In addition, Table IV shows that all HTMT values were less than 0.85 [85], confirming that the discriminant validity of the measurement model was satisfactory. The structural model is shown in Fig. 2.

TABLE I. Descriptive Analysis of Demographic Variables

Variable	Item	Frequency	Percentage %
Gender	Male	87	34.4
	Female	166	65.6
Age	20-29	23	9.1
	30-39	176	69.6
	40-49	37	14.6
	50 and above	17	6.7
Organization	Ministries	210	83
	Departments	10	4

	Agencies	33	13
Job schemes	Information Technology Officer	74	29.2
	Administrative Officer	112	44.3
	Finance Officer	9	3.6
	Human Resources	33	13
	Archive Officer	21	8.3
	Engineer	4	1.6
Grade	14	11	4.3
	13	17	6.7
	12	48	19
	10	73	28.9
	9	104	41.1
Years of service	1-5	22	8.7
	6-10	105	41.5
	11-15	83	32.8
	16-20	15	5.9
	More than 20 years	26	10.3
Roles in ERM	System Administrator	9	3.6
	Information Records Manager	43	17
	Support	14	5.5
	End User	187	73.95

TABLE II. MEASUREMENT MODEL ANALYSIS

Construct	Indicator	Loadings	CR	AVE	Discriminant Validity
PE	PE1	0.841	0.962	0.834	Yes
	PE2	0.809			
	PE3	0.851			
	PE4	0.711			
	PE5	0.829			
EE	EE1	0.917	0.922	0.747	Yes
	EE2	0.775			
	EE3	0.903			
	EE4	0.877			
	EE5	0.836			
SI	SI1	0.803	0.956	0.845	Yes
	SI2	0.853			
	SI3	0.921			
	SI4	0.900			
	SI5	0.916			
FC	FC1	0.891	0.905	0.656	Yes
	FC2	0.883			
	FC3	0.908			
	FC4	0.860			
	FC5	0.923			
ATU	ATU1	0.917	0.936	0.745	Yes
	ATU2	0.901			
	ATU3	0.938			
	ATU4	0.896			
	ATU5	0.912			
I	I1	0.895	0.974	0.902	Yes
	I2	0.943			
	I3	0.927			

	I4	0.912			
D	D1	0.823	0.952	0.798	Yes
	D2	0.884			
	D3	0.878			
	D4	0.872			
K	K1	0.921	0.953	0.802	Yes
	K2	0.947			
	K3	0.941			
	K4	0.941			
SK	SK1	0.932	0.945	0.774	Yes
	SK2	0.939			
	SK3	0.932			
	SK4	0.945			
NTIM	NTIM1	0.854	0.975	0.906	Yes
	NTIM2	0.916			
	NTIM3	0.948			
	NTIM4	0.907			
	NTIM5	0.849			
TLP	TLP1	0.888	0.947	0.818	Yes
	TLP2	0.906			
	TLP3	0.926			
	TLP4	0.899			

TABLE III. FORNELL-LARCKER CRITERION

	ATU	D	I	PE	EE	K	FC	NTIM	SI	SK	TLP
ATU	**0.913**										
D	0.205	**0.864**									
I	0.120	0.394	**0.919**								
PE	0.253	0.613	0.357	**0.810**							
EE	0.231	0.522	0.215	0.645	**0.863**						
K	0.236	0.639	0.336	0.503	0.646	**0.950**					
FC	0.182	0.617	0.355	0.741	0.598	0.483	**0.893**				
NTIM	0.116	0.330	0.178	0.675	0.630	0.295	0.638	**0.896**			
SI	0.276	0.571	0.334	0.661	0.659	0.541	0.656	0.578	**0.880**		
SK	0.377	0.585	0.252	0.578	0.412	0.347	0.554	0.460	0.455	**0.952**	
TLP	0.164	0.286	0.117	0.627	0.501	0.200	0.521	0.724	0.522	0.386	**0.905**

TABLE IV. HETEROTRAIT-MONOTRAIT RATIO (HTMT)

	ATU	D	I	PE	EE	K	FC	NTIM	SI	SK	TLP
ATU											
D	**0.190**										
I	0.134	**0.417**									
PE	0.249	0.694	**0.387**								
EE	0.217	0.575	0.222	**0.727**							
K	0.231	0.687	0.342	0.550	**0.688**						
FC	0.150	0.667	0.372	0.821	0.644	**0.507**					
NTIM	0.093	0.359	0.182	0.747	0.680	0.310	**0.677**				
SI	0.264	0.613	0.342	0.725	0.709	0.565	0.690	**0.613**			
SK	0.363	0.632	0.273	0.633	0.440	0.361	0.579	0.482	**0.469**		
TLP	0.151	0.308	0.136	0.701	0.544	0.211	0.557	0.777	0.557	**1.**	

C. Structural Model Analysis

Once the measurement model meets the reliability and validity criteria, the next evaluation is a structured model to assess the strength of the relationships among the constructs and the model's predictive ability. This evaluation involves collinearity analysis, hypothesis testing, coefficient of determination (R^2), effect size (f^2), and predictive relevance (Q^2) using SmartPLS software.

1) Collinearity assessment: Refers to high correlation between independent variables, which can affect the accuracy of the regression coefficients. The variance inflation factor (VIF) is used to check for collinearity. A VIF value of 5 or higher suggests possible collinearity, while a value below 5 indicates no issues [84]. In this analysis, all independent variables had VIF values below the threshold, as shown in Table V, meaning no collinearity problems in the study model.

2) Hypothesis testing (path coefficient): Conducted through bootstrap analysis with 500 random resamples. Table VI summarizes the hypothesis testing and path coefficients. The relationship between constructs was assessed using path coefficients (β), t-values (>1.65), and p-values (<0.05) in a one-tailed test. The results showed that nine hypotheses were supported with significant, positive relationships, while one hypothesis (h6: i → ntim) was rejected because the relationship was not significant and negative. The majority of relationships were consistent with the study hypothesis.

3) Coefficient of determination (r^2): The r^2 value in Table VII indicates the level of variance of the dependent construct that the independent construct can explain. The analysis found that the r^2 values for ntim and tlp were 0.633 and 0.524, respectively, both at a moderate level [84]. This means that the independent construct explained 63.3% of ntim variance, while ntim explained 52.4% of tlp variance, indicating that the model's predictive ability was moderate.

4) Effect size (f^2): The effect size (f^2) was used to assess the contribution of each independent construct to the predictive accuracy of the model. The results in Table VIII showed that most constructs had small effects on ntim, while the ntim → tlp relationship showed a large effect (f^2 =1.099). Construct: I did not show any effect, consistent with the non-significant path coefficient results. The majority of relationships in the model showed small effects, with one main relationship at a large level.

5) Predictive relevance (q^2): Predictive relevance was assessed using the Stone–Geisser (q^2) statistic via the plspredict procedure. Table IX shows the predictive relevance (q2) values. A q^2 value greater than zero indicates that the model has good predictive ability. The results show that the q^2 values for ntim (0.561) and tlp (0.410) are high, indicating that the study model has high predictive relevance.

TABLE V. CONSTRUCT COLLINEARITY ASSESSMENT (VIF)

Dependent constructs	Independent constructs	VIF
NTIM	ATU	1.229
	D	2.613
	I	1.269
	PE	2.979
	EE	2.584
	K	2.309
	FC	2.309
	SI	2.414
	SK	1.976
TLP	NTIM	1

TABLE VI. PATH COEFFICIENT TEST

Hypothesis	Construct relationship	β	t-value	p-value	Discussion
H1	PE -> NTIM	0.343	*3.065	*0.002	Supported
H2	EE -> NTIM	0.375	*3.187	*0.002	Supported
H3	SI -> NTIM	0.154	*2.494	*0.013	Supported
H4	FC -> NTIM	0.244	*2.061	*0.040	Supported
H5	ATU -> NTIM	0.108	*2.551	*0.011	Supported
H6	I -> NTIM	-0.030	0.450	0.653	Rejected
H7	D -> NTIM	0.268	*2.883	*0.004	Supported
H8	K -> NTIM	0.103	*2.339	*0.020	Supported
H9	SK -> NTIM	0.167	*2.442	*0.015	Supported
H10	NTIM -> TLP	0.724	*12.785	*0.000	Supported

TABLE VII. ANALYSIS OF MODEL PREDICTION ACCURACY BASED ON R^2 VALUE

Dependent constructs	R^2 Value
NTIM	0.633 (Moderate)
TLP	0.524 (Moderate)

TABLE VIII. RESULTS OF THE ANALYSIS OF THE EFFECT SIZE VALUE (F^2)

Dependent constructs	Independent constructs	Effect size (f^2)	Impact of change
NTIM	ATU	0.026	Small
	D	0.075	Small
	I	*0.002	No effect
	PE	0.108	Small
	EE	0.149	Small
	K	0.035	Small
	FC	0.059	Small
	SI	0.027	Small
	SK	0.038	Small
TLP	NTIM	1.099	Big

TABLE IX. RESULTS OF THE ANALYSIS OF THE EFFECT SIZE VALUE (Q^2)

Dependent constructs	R^2 Value	Q^2 Value
NTIM	0.633	0.561
TLP	0.524	0.410

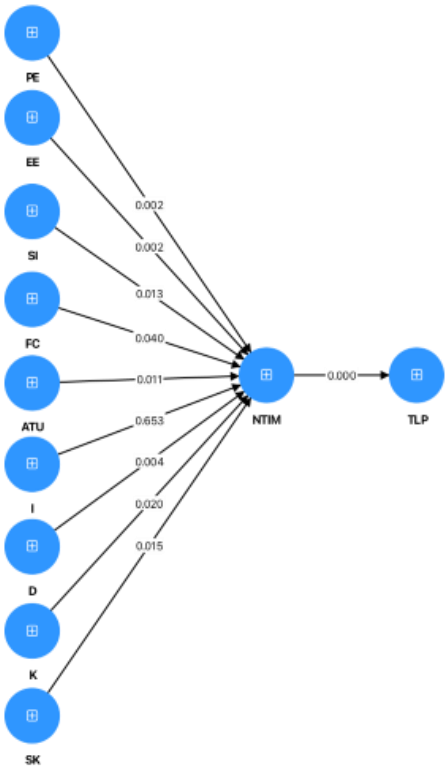


Fig. 2. Structural model.

V. CONCLUSION AND DISCUSSION

A. Hypothesis Testing Results

The results of the analysis showed that eight out of nine predictor factors had a positive and significant relationship with behavioral intention to adopt blockchain technology (NTIM), namely Performance Expectations (H1), Effort Expectations (H2), Social Influence (H3), Facilitating Conditions (H4), Attitudes Towards Technology (H5), Policies (H7), Security (H8), and Leadership Support (H9). On the other hand, Infrastructure Quality (H6) was found to be insignificant.

1) Individual factors: Performance expectations ($\beta = 0.343$, $P < 0.01$) and effort expectations ($\beta = 0.375$, $P < 0.01$) had a

significant positive effect on adoption intention. This shows that perceptions of technology's benefits and ease of use are the main drivers of user intention, in line with the UTAUT theory [57]. In addition, attitude towards technology also significantly influenced intention, although with a smaller effect, indicating the importance of user openness to innovation.

2) Organisational and environmental factors: Social influence ($\beta = 0.154$, $P < 0.05$), facilitating conditions ($\beta = 0.244$, $P < 0.05$), Policy ($\beta = 0.268$, $P < 0.01$), security ($\beta = 0.103$, $P < 0.05$), and leadership support ($\beta = 0.167$, $P < 0.05$) were found to be significant in shaping adoption intentions. This finding underscores that organizational support, policy clarity, confidence in information security, and leadership commitment are important in a structured, rule-based public sector.

3) Infrastructure quality: Hypothesis H6 regarding infrastructure quality was not supported ($\beta = -0.030$, $P > 0.05$). This finding suggests that respondents considered technological infrastructure sufficient and that it is no longer a major determinant of intention formation. Instead, human and organizational factors are more dominant in influencing blockchain technology adoption.

4) Relationship between intention and usage behavior: The results of the analysis showed a very strong and significant relationship between behavioral intention (NTIM) and actual usage behavior (TLP) ($\beta = 0.724$, $P < 0.001$). This finding confirms that intention is a key predictor of actual technology usage, in line with TAM and UTAUT theories [49, 57].

B. Key Findings

The research model showed good explanatory power, with R^2 values of 0.633 for behavioral intention and 0.524 for usage behavior. The predictive analysis also showed a positive Q^2 value, indicating moderate to good predictive power. The effect size analysis showed that the NTIM → TLP relationship had a large effect ($f^2 = 1.099$), while other factors on intention showed small to medium effects. The study also found that people's intentions to adopt blockchain technology depend on individual, organizational, and environmental factors. Performance expectations, effort expectations, social influence, facilitating conditions, attitudes toward technology, policy, security, and leadership support all played important roles, but infrastructure quality did not have a significant impact. The link between intentions and actual use shows that intentions are key before people start using new technology. Even with good organizational and infrastructure support, adoption will not happen unless users have strong intentions. This supports the TAM and UTAUT theories for the Malaysian public sector ERM. The integration of TAM and UTAUT was theoretically motivated by the complementary strengths of both models. TAM examines user perceptions of usefulness and ease of use, while UTAUT captures broader social and facilitating influences. Rather than treating them as competing models, this study adopts an integrative approach suited to complex public-sector environments, where both individual beliefs and institutional conditions influence adoption decisions. The model's relatively strong explanatory power ($R^2 = 0.633$) suggests that the combined framework provides robust predictive capability. Future studies may conduct nested model comparisons to examine the incremental predictive contributions of each framework. In addition to R^2, predictive relevance (Q^2) and effect size (f^2) were assessed to provide a comprehensive PLS-SEM evaluation. The Q^2 values for behavioral intention (0.561) and usage behavior (0.410) were above zero, indicating strong predictive relevance. Effect size analysis showed that behavioral intention had a large effect on usage behavior ($f^2 = 1.099$), while other predictors demonstrated small to moderate effects.

1) Theoretical implications: The findings of this study confirm the relevance of TAM and UTAUT theories in explaining blockchain technology adoption in the public sector. The main construct of UTAUT shows a significant relationship with behavioral intention, and behavioral intention is a strong predictor of actual usage behavior. The moderate to high R^2 value and large F^2 effect size for the intention–usage relationship support the view that technology adoption is a gradual process driven by intention formation. The lack of significance of infrastructure quality suggests the need to adapt the technology adoption model to the local context, where policy, security, and leadership factors are more dominant in the public sector.

2) Practical implications: The findings highlight the importance of clear company rules, strong leadership, and a focus on keeping information secure to encourage the adoption of blockchain technology. Also, providing thorough training, good ways to share information, and designs that focus on users are needed to set clear goals for how well the technology should work and how much effort it will take, and to help people feel good about using it. Evenhough the quality of basic infrastructure is not the main issue, having good technical support and using resources effectively are still very important for making the technology work.

VI. Limitations and Future Work

Although this study followed a systematic approach and relied on a solid theoretical framework, it has some limitations. The research used a cross-sectional survey, collecting data at just one point in time. As a result, the findings show only respondents' views on adopting blockchain technology in ERM during the study period and do not capture how user attitudes or behaviors might change over time. The study only included respondents from the public sector in Malaysia. This means the results mainly apply to public-sector organizations in that country and may not fully reflect the private sector or organizations in other countries with different laws, cultures, or levels of technology. This study also used only quantitative methods through a questionnaire. While this works well for testing relationships in the model, it does not allow for a deeper look at the context, practical challenges, or real-world issues of using blockchain technology that could affect adoption in public sector organizations. Additionally, the study examines certain variables based on technology adoption models such as TAM and UTAUT. While these models are well supported by research, other factors such as legal issues, costs, infrastructure

readiness, and data governance were not directly tested in this study.

Future research should include a longitudinal study to track changes in perceptions, behavioral intentions, and actual use of blockchain technology in ERM. This would help researchers understand how the technology adoption process develops over time. Research could also include private-sector organizations and quasi-governmental agencies, or compare results across countries that may show significant differences in the factors affecting the adoption of blockchain technology in ERM. A mixed methods approach could combine in-depth interviews, case studies, or document analysis. This can add valuable qualitative insights to support quantitative results, especially regarding operational-level challenges. Future research can also include variables such as information technology governance, organizational readiness, trust, technology costs and risks, and compliance with policies and laws, to expand the conceptual model. These variables may enhance the model's predictive ability and the relevance of the findings to policymakers and practitioners. Finally, experimental or pilot studies on blockchain-based electronic records management systems in public-sector organizations would be valuable. These studies could confirm current findings and help develop practical guidelines and policies for using blockchain in public records and archives management.

ACKNOWLEDGMENT

This study was funded by GUP-2024-045.

DECLARATION ON THE USE OF GENERATIVE AI

The authors acknowledge the use of a generative AI tool to enhance language clarity and assist in correct grammar. However, all scientific ideas, conceptual design, interpretations, and conclusions were independently developed by the authors. All AI-supported content was carefully revised, improved, and validated by authors, and they accept full responsibility for the manuscript's quality and integrity.

REFERENCES

[1] N. H. Ahmad Tajudin and N. S. Ahmad Shukor, "Conceptualising effective record management of Malaysian public sector", business review journal, vol. 9, no. 2, pp. 157–170, Dec. 2024.

[2] D. Yuqiao and S. Linyu "Exploring the mechanisms for records management's digital transformation: a case study from China.", records management journal, vol. 31, pp. 34-47, 2020.

[3] Malaysia: Personal data protection (amendment) bill 2024, July 2024.

[4] American Library Association. "Guidelines: competencies for special collections professionals." Last modified July 8, 2008.

[5] N. Saadan, K.F Khairi, A. Ahmad, and A.M. Tahir, "Mapping blockchain adoption challenges: a bibliometric analysis," the journal of muamalat and islamic finance research, pp.101-126, 2024.

[6] Hossain MI, Steigner T, Hussain MI, Akther A. "Enhancing data integrity and traceability in industry cyber physical systems (ICPS) through Blockchain technology: A comprehensive approach," arXiv preprint arXiv:2405.04837, May 8, 2024.

[7] Z. Ahmad and S. Abdullah, "Blockchain Technology Adoption in Malaysia's e-Government Services: Insights from Case Studies and a Framework for Implementation", Int. J. of Adv. Sci. Comp. and Eng., vol. 7, no. 2, pp. 36–41, Aug. 2025.

[8] N. Saadan and K. F. Khairi, "Exploring the Suitability of Blockchain Technology in Improving Inter-Organizational Collaboration Workflows in E-Payment Processes at the Transport License Department of Malaysia", *IMBR*, vol. 17, no. 3(I)S, pp. 102-114, Oct. 2025.

[9] A. A. Aziz, Z. M. Yusof, U. A. Mokhtar, and D. I. Jambari, "A Conceptual Model for Electronic Document and Records Management System Adoption in Malaysian Public Sector", Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 4, pp. 1191–1197, Aug. 2018.

[10] R. A. Yaacob and R. M. Sabai, "Electronic records management in Malaysia: A case study in one government agency," in Proc. Asia-Pacific Conf. Library & Information Education & Practice (A-LIEP2011), 2011.

[11] S. K. Tulli, "The unified theory of acceptance and use of technology (UTAUT) model in evaluating NetSuite ERP adoption,", vol. 3, no. 1, 2024.

[12] J. S. Heier et al., "Efficacy, durability, and safety of intravitreal faricimab up to every 16 weeks for neovascular age-related macular degeneration (TENAYA and LUCERNE): Two randomised, double-masked, phase 3 trials," Lancet, vol. 399, no. 10326, pp. 729–740, Feb. 2022.

[13] B. Wiggins, "Records and information management," Int. J. Information Management, vol. 33, pp. 976–977, 2013.

[14] G. Yeo, "Records and information management," Archives and Records, vol. 40, no. 3, pp. 305–308, 2019.

[15] Y. Raja Abdullah and S. Roseline, "Electronic records management in Malaysia: A case study in one government agency," in Asia Pacific Conf. Library & Information Education & Practice, pp. 420–433, 2011.

[16] W. Saffady, Records and Information Management. London: Rowman & Littlefield, 2021.

[17] A. M. Omar and M. O. Abdullahi, "A bibliometric analysis of sustainable digital transformation in developing countries' higher education," Front. Educ., vol. 9, 2024.

[18] N. Su, F. Jia, S. L. Pan, and C. Hsu, "ICTs and rural e-governance: From digital design to public policy," Information Systems Journal, 2025.

[19] Maintec Technologies, "Report," 2023.

[20] S. Samsudin and M. H. Hanafiah, "Challenges in applying IR4.0 technologies to improve organizational structure and positions: The case of Malaysian public agencies," Jurnal Pengurusan, vol. 70, pp. 1–15, 2024.

[21] K. L. Newray, "Improving public record management system at government ministries and agencies: Factors to consider and action steps to take," Int. J. Multidisciplinary Research, vol. 5, no. 6, 2023.

[22] M. Kassen, "Blockchain and e-government innovation: Automation of public information processes," Information Systems, vol. 103, Art. no. 101862, 2022.

[23] Arkib Negara Malaysia (ANM), "Report," 2023.

[24] N. H. A. Tajudin and N. S. A. Shukor, "Conceptualising effective record management of Malaysian public sector," Selangor Business Review, pp. 157–170, 2024.

[25] J. Babikian, "Balancing acts: Ethics, regulation, and accountability in AI law and policy," J. Management Science Research Review, vol. 1, no. 1, pp. 1–23, 2023.

[26] L. Qudus, "Cybersecurity governance: Strengthening policy frameworks to address global cybercrime and data privacy challenges," Int. J. Science and Research Archive, vol. 14, no. 1, pp. 1146–1163, 2025.

[27] P. K. Pemmasani and M. A. A. Nasaruddin, "Resilient IT strategies for governmental disaster response and crisis management," Int. J. Acta Informatica, vol. 1, no. 1, pp. 151–163, 2022.

[28] K. Al-Assaf, Z. Bahroun, and V. Ahmed, "Transforming service quality in healthcare: A comprehensive review of healthcare 4.0 and its impact," Informatics, vol. 11, no. 4, 2024.

[29] S. K. S. A. Mutalib, Email record keeping in the government sector: A case study of Malaysia. PhD dissertation, Univ. Northumbria, Newcastle, 2018.

[30] A. S. George, "Digital hoarding: The rising environmental and personal costs of information overload," Partners Universal Multidisciplinary Research Journal, vol. 1, no. 2, pp. 51–67, 2024.

[31] B. Singh, "Cyber security for databases: Advanced strategies for threat detection and response," Int. J. Current Engineering and Scientific Research, vol. 2.

[32] Ö. Aslan et al., "A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions," Electronics, vol. 12, no. 6, Art. no. 1333, 2023.

[33] K. Thibodeau, "Overview of technological approaches to digital preservation and challenges in coming years," 2024.

[34] J. Zhang and Z. Chen, "Exploring human resource management digital transformation in the digital age," J. Knowledge Economy, vol. 15, no. 1, pp. 1482–1498, 2024.

[35] A. J. Akbarfam et al., "ForensiBlock: A provenance-driven blockchain framework for data forensics and auditability," arXiv preprint, 2023.

[36] U. S. Ugbaja et al., "Conceptual framework for role-based network access management to minimize unauthorized data exposure," Int. J. Social Science Exceptional Research, vol. 2, no. 1, pp. 211–221, 2023.

[37] M. Kapanda, An investigation of the impact on electronic records management on service delivery in the department of immigration in Zambia. PhD dissertation, Univ. Zambia, 2024.

[38] N. W. D. Jayathilake and S. M. C. Seneviratne, "The investigation of the awareness of implementing blockchain technology in audit trails among auditors," J. Accounting Research, Organization and Economics, vol. 5, no. 2, pp. 109–123, 2022.

[39] M. M. Amalia, "Enhancing accountability and transparency in the public sector: A comprehensive review," The ES Accounting and Finance, vol. 1, no. 3, pp. 160–168, 2023.

[40] "Trusting records: Is blockchain technology the answer?," Records Management Journal, vol. 26, no. 2, pp. 110–139, 2016.

[41] T. Miller, Decentralising trust and information symmetry: Seeking retrospective common knowledge in blockchain enabled supply chains. PhD dissertation, Queensland Univ. Technology, 2025.

[42] J. Xiao et al., "A collaborative auditing scheme with dynamic data updates based on blockchain," Connection Science, vol. 35, no. 1, Art. no. 2213863, 2023.

[43] E. Tan, S. Mahula, and J. Crompvoets, "Blockchain governance in the public sector: A conceptual framework for public management," Government Information Quarterly, vol. 39, no. 1, Art. no. 101625, 2022.

[44] M. S. Kiu et al., "Blockchain integration into electronic document management (EDM) system in construction common data environment," Smart and Sustainable Built Environment, vol. 13, no. 1, pp. 117–132, 2024.

[45] V. L. Lemieux, Searching for trust: Blockchain technology in an age of disinformation. Cambridge: Cambridge Univ. Press, 2022.

[46] N. Marangunić and A. Granić, "Technology acceptance model: A literature review from 1986 to 2013," Universal Access in the Information Society, vol. 14, no. 1, pp. 81–95, 2015.

[47] P. Surendran, "Technology acceptance model: A survey of literature," Int. J. Business and Social Research, vol. 2, no. 4, pp. 175–178, 2012.

[48] R. Ibrahim et al., "E-learning acceptance based on technology acceptance model (TAM)," J. Fundamental and Applied Sciences, vol. 9, no. 4S, pp. 871–889, 2017.

[49] M. Muchran and A. S. Ahmar, "Application of TAM model to the use of information technology," arXiv preprint, 2019, arXiv:1901.11358.

[50] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," Management Science, vol. 35, no. 8, pp. 982–1003, 1989, doi: 10.1287/mnsc.35.8.982.

[51] S. Nagy and N. Hajdu, "Consumer acceptance of the use of artificial intelligence in online shopping: Evidence from Hungary," arXiv preprint, 2022.

[52] M. Al-Emran, "Beyond technology acceptance: Development and evaluation of technology-environmental, economic, and social sustainability theory," Technology in Society, vol. 75, Art. no. 102383, 2023.

[53] M. M. Cioc et al., "Behavioral intentions to use energy efficiency smart solutions under the impact of social influence: An extended TAM approach," Applied Sciences, vol. 13, no. 18, Art. no. 1024, 2023.

[54] A. V. Martín-García, F. Martínez-Abad, and D. Reyes-González, "TAM and stages of adoption of blended learning in higher education," British Journal of Educational Technology, vol. 50, no. 5, pp. 2484–2500, 2019.

[55] W. K. Tse, X. Dai, Y. M. Lee, and D. Lu, "User acceptance of blockchain technology in financial applications: Information security, technology awareness and privacy aspects," Blockchains, vol. 2, no. 3, pp. 299–311, 2024, doi: 10.3390/blockchains2030014.

[56] "Factors influencing behavioral intention to use e-learning in higher education during the COVID-19 pandemic: A meta-analytic review based on the UTAUT2 model," Education and Information Technologies, vol. 30, 2025.

[57] V. Venkatesh et al., "User acceptance of information technology: Toward a unified view," MIS Quarterly, pp. 425–478, 2003.

[58] D. Marikyan, S. Papagiannidis, and G. Stewart, "Technology acceptance research: Meta-analysis," Journal of Information Science, 2023.

[59] Y. Jadil, N. P. Rana, and Y. K. Dwivedi, "A meta-analysis of the UTAUT model in the mobile banking literature," Journal of Business Research, vol. 132, pp. 354–372, 2021.

[60] Y. K. Dwivedi et al., "A meta-analysis of the unified theory of acceptance and use of technology (UTAUT)," in Proc. IFIP Int. Working Conf. Governance and Sustainability in Information Systems, Berlin, Heidelberg: Springer, 2011.

[61] C. Martins, T. Oliveira, and A. Popovič, "Understanding the internet banking adoption: A unified theory of acceptance and use of technology and perceived risk application," Int. J. Information Management, vol. 34, no. 1, pp. 1–13, 2014.

[62] E. L. Slade et al., "Modeling consumers' adoption intentions of remote mobile payments in the United Kingdom: Extending UTAUT with innovativeness, risk, and trust," Psychology & Marketing, vol. 32, no. 8, pp. 860–873, 2015.

[63] Technology acceptance models in health informatics: TAM and UTAUT," Int. J. Medical Informatics, vol. 129, pp. 13–23, 2019.

[64] [5] A. K. Shrestha and J. Vassileva, "User acceptance of usable blockchain-based research data sharing system: An extended TAM-based study," in Proc. IEEE Int. Conf. Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), 2019.

[65] [6] T. Oliveira et al., "Extending the understanding of mobile banking adoption: When UTAUT meets TTF and ITM," Int. J. Information Management, vol. 34, no. 5, pp. 689–703, 2014.

[66] N. P. Rana et al., "Citizens' adoption of an electronic government system: Towards a unified view," Information Systems Frontiers, vol. 19, no. 3, pp. 549–568, 2017.

[67] S.-Y. Hung et al., "User acceptance of intergovernmental services: An example of electronic document management system," Government Information Quarterly, vol. 26, no. 2, pp. 387–397, 2009.

[68] A. K. Shrestha and J. Vassileva, "User acceptance of usable blockchain-based research data sharing system: An extended TAM-based study," arXiv preprint, 2019.

[69] M. Kouhizadeh, S. Saberi, and J. Sarkis, "Blockchain technology and the sustainable supply chain: Theoretically exploring adoption barriers," Int. J. Production Economics, vol. 231, 2021.

[70] D. Gefen, E. Karahanna, and D. W. Straub, "Trust and TAM in online shopping: An integrated model," MIS Quarterly, pp. 51–90, 2003.

[71] M. M. Queiroz and S. F. Wamba, "Blockchain adoption challenges in supply chain: An empirical investigation of the main drivers in India and the USA," Int. J. Information Management, vol. 46, pp. 70–82, 2019.

[72] A. Nikiforova et al., "Responsible AI adoption in the public sector: A data-centric taxonomy of AI adoption challenges," arXiv preprint, 2025.

[73] M. D. Williams, N. P. Rana, and Y. K. Dwivedi, "The unified theory of acceptance and use of technology (UTAUT): A literature review," J. Enterprise Information Management, vol. 28, no. 3, pp. 443–488, 2015.

[74] M. R. Albous, A. AlKandari, and A. L. Anouze, "From vision to validation: A theory- and data-driven construction of a GCC-specific AI adoption index," arXiv preprint, 2025.

[75] J. Leng et al., "Blockchain security: A survey of techniques and research directions," IEEE Trans. Services Computing, vol. 15, no. 4, pp. 2490–2510, 2020.

[76] E. Toufaily, T. Zalan, and S. B. Dhaou, "A framework of blockchain technology adoption: An investigation of challenges and expected value," Information & Management, vol. 58, no. 3, 2021.

[77] S. Salem and N. Ali, "A proposed adoption model for blockchain technology using the unified theory of acceptance and use of technology (UTAUT)," Open Int. J. Informatics, 2019.

[78] M. Janssen et al., "Data governance: Organizing data for trustworthy artificial intelligence," Government Information Quarterly, vol. 37, no. 3, 2020.

[79] L. Hughes et al., "Blockchain research, practice and policy: Applications, benefits, limitations, emerging research themes and research agenda," Int. J. Information Management, vol. 49, pp. 114–129, 2019.

[80] L. Manny et al., "Barriers to the digital transformation of infrastructure sectors," Policy Sciences, vol. 54, no. 4, pp. 943–983, 2021.

[81] I. Ajzen, "The theory of planned behavior," Organizational Behavior and Human Decision Processes, vol. 50, no. 2, pp. 179–211, 1991.

[82] M. I. Latif and Z. Zakaria, "Factors determine the behavioural intention in adopting blockchain technology by Malaysian public sector officers," J. Advanced Research in Business and Management Studies, vol. 20, no. 1, pp. 34–43, 2020.

[83] D. D. Nulty, "The adequacy of response rates to online and paper surveys: What can be done?," Assessment & Evaluation in Higher Education, vol. 33, no. 3, pp. 301–314, 2008.

[84] [4] J. F. Hair Jr. et al., "PLS-SEM or CB-SEM: Updated guidelines on which method to use," Int. J. Multivariate Data Analysis, vol. 1, no. 2, pp. 107–123, 2017.

[85] J. Henseler, C. M. Ringle, and R. R. Sinkovics, "The use of partial least squares path modeling in international marketing," in New Challenges to International Marketing. Emerald Group Publishing, 2009, pp. 277–319.

[86] W. W. Chin, "Commentary: Issues and opinion on structural equation modeling," MIS Quarterly, pp. vii–xvi, 1998.

[87] N. Urbach and F. Ahlemann, "Structural equation modeling in information systems research using partial least squares," J. Information Technology Theory and Application, vol. 11, no. 2, p. 2, 2010.

[88] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," J. Marketing Research, vol. 18, no. 1, pp. 39–50, 1981.

Reproducible Prediction Framework of Customer Churn Using Machine Learning, Advanced Data Science and Business Intelligence Techniques

Younes KOULOU, Norelislam EL HAMI
Science and Engineering Laboratory-ENSA, Ibn Tofail University, Kenitra, Morocco

***Abstract*—The telecommunications sector has evolved in recent years, resulting in intense competition and high customer acquisition costs. As a result, retaining customers has become a key concern for telecom operators. In this work, we propose the design and implementation of a complete customer churn prediction system that combines data science, machine learning and business intelligence approaches. The methodology is structured into five main steps: exploratory data analysis, development of an ETL pipeline, feature engineering, predictive modeling using a Random Forest algorithm, and the creation of decision-support dashboards in Power BI. Random Forest demonstrated higher performance with AUC-ROC of 0,85 and the results demonstrated that the main predictors of churn are monthly charges, contract type, and customer tenure. Our approach, which validated by a confusion matrix, offers decision-makers an operational tool to anticipate departures and implement targeted loyalty actions. This study proposes a reproducible methodological framework for companies facing the problem of churn and contributes to the use of machine learning in relationship marketing.**

***Keywords*—Churn prediction; telecommunications; machine learning; random forest; business intelligence; ETL; Power BI; feature engineering; decision support**

I. Introduction

In today's digital economy, the telecommunications sector is characterized by intense competition, saturated mature markets, and particularly high customer acquisition costs. Acquiring a new customer can cost up to 25 times more than retaining an existing one [1]. In this context, customer churn poses a direct threat to the profitability and long-term viability of businesses.

For this reason, the phenomenon of churn takes on particular strategic importance in telecommunications where the sudden interruption of contractual revenues significantly affects financial forecasts, loyal customers tend to subscribe to more additional services, and the quality of customer relationship represents a decisive competitive advantage in such a market where differentiation by supply becomes difficult [2].

Churn prediction has been addressed from various methodological perspectives. There are ensemble learning approaches, which currently dominate, utilize algorithms such as Decision Trees, Random Forest, XGBoost, LightGBM, and CatBoost. These algorithms demonstrate superior performance for churn prediction in the telecommunications sector [3] [4]. In this context, [4] explored the effectiveness of ensemble learning techniques by combining several basic learners, such as decision trees and gradient boosting in an ensemble framework in order to improve predictive accuracy to that of individual learners [5]. In parallel, the advent of Deep Learning has opened up new perspectives. While other researchers have focused on the interpretability of models and the identification of critical factors affecting churn [6].

A recent systematic review analyzed 240 studies, published between 2020 and 2024, on churn prediction. The authors identified several persistent challenges: class imbalance, model interpretability, the concept of drift such as the temporal evolution of customer behavior, and the limited use of profit-oriented metrics [4].

Unlike most existing studies, which focus on algorithmic performance, we propose a holistic approach integrating the entire data value chain within a comprehensive methodological framework that includes: a rigorous ETL (Extract, Transform, Load) process containing business rules; a feature engineering phase to create interpretable variables; Optimized modeling with cross-validation and hyperparameter search; an interactive application for Operational deployment; and integration of business intelligence via BI dashboards.

This approach addresses the need to bring predictive models closer to the operational needs of businesses.

The structure of this article is as follows: Section 2 details our methodology. Section 3 presents the results. And Section 4 discusses the results and suggests some perspectives.

II. Related Work

Numerous studies have been conducted to predict customer churn. Despite their interpretability, early approaches lacked the ability to capture complex non-linear relationships through traditional statistical methods like logistic regression. The introduction of machine learning has resulted in a significant improvement in predictive performance.

Churn prediction is now dominated by ensemble methods. According to [3], the authors demonstrated that combining multiple base learners (decision trees, gradient boosting) in an ensemble framework can result in superior predictive accuracy than individual learners. Telecommunications data [3] [4] has demonstrated robust performance for Random Forest, XGBoost, LightGBM, and CatBoost.

New possibilities have been opened up by deep learning architectures simultaneously. The author in [4] conducted a

systematic review examining 240 publications from 2020 to 2024 and found that the use of temporal dependencies-captured LSTM networks is gradually being applied to churn prediction. Data volume requirements and interpretability concerns limit their implementation.

Interpretability of models has become a crucial concern. According to the authors of [6], XAI-Churn TriBoost is a model that is interpretable and integrates XGBoost, CatBoost, and LightGBM into a weighted voting ensemble, and it combines LIME and SHAP techniques to explain individual predictions and uncover critical factors impacting churn.

According to the systematic review by [4], there are many challenges that persist: class imbalance, model interpretability, concept drift (temporal evolution of customer behavior), and the limited use of profit-oriented metrics. The primary focus of existing studies is on algorithmic performance, leaving out the operational integration of predictive models into business processes. Few works address the complete data value chain from extraction to strategic visualization, which constitutes the main contribution of our research.

III. Methodology

The proposed approach is broken down into five main phases, as illustrated in Fig. 1.

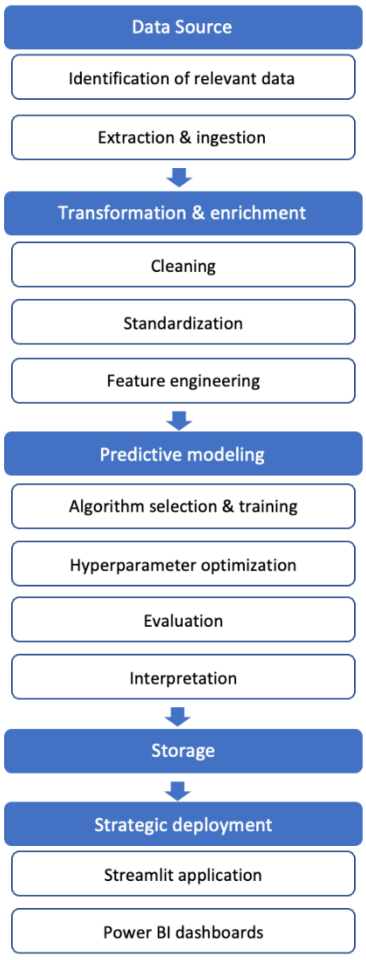


Fig. 1. Methodological diagram.

A. Data Used

The dataset used in this study comes from a telecommunications company that provides phone and internet services. It comprises 7043 observations and 21 variables. Table I presents the main characteristics of this Dataset.

TABLE I. Dataset Characteristics

Characteristic	Value
Total observations	7 043
Number of features	21
Non-churn customers	5 174 (73,5%)
Churn customers	1 869 (26,5%)
Class ratio (non-churn:churn)	2.8:1
Memory size	7.79 MB

B. Data Extraction and Ingestion

This initial phase consists of loading the data from its source and verifying its structural integrity. The dataset's informative diversity covers four categories of variables (see Table II):

TABLE II. Typology of Dataset Variables

Category	Variables	Description
Demographics	Gender, Senior Citizen, Partner, Dependents	Sociodemographic profile
Services	Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, StreamingTV, Streaming Movies	Package of subscribed services
Contractual	Contract, Paperless Billing, Payment Method	Contractual terms
Financial	Monthly Charges, Total Charges, Tenure	Financial indicators and loyalty
Target	Churn	Predictable variable (Yes/No)

C. Pipeline ETL and Feature Engineering

The reliability of predictive models depends largely on the quality of the input data. In this step, we implemented an ETL pipeline to progressively clean, structure, and transform the data as follows:

1) Data cleaning: This step aims to detect and correct data quality issues [7]. During the initial analysis, the totalcharges variable was found to contain 11 observations with spaces in place of numerical values. After examination, these cases corresponded to customers with zero tenure. To resolve this issue, the variable was converted to a numeric type, where missing values are automatically treated as null and then imputed as 0. This solution preserved observations while maintaining data integrity.

2) Handling of categorical variables: This step aims to reduce inconsistencies in categorical variables by grouping similar categories under a unified representation [8]. For example, We found that the service variables had three categories: "yes", "no", and variants such as "no internet

service" and "no telephone service". To harmonize the data, we grouped these variants into a single category "no".

3) Feature engineering: The goal of this step is to enrich the dataset by creating derived variables that offer greater analytical value and contribute to improved predictive performance. Eight new variables were created:

- ChurnFlag: conversion of the Churn variable into a binary format.
- Tenure Group: Segmentation of tenure into four classes: 0-1 year, 1-2 years, 2-4 years, >4 years.
- Total Services: This represents the number of subscribed services by each customer. Possible values 0 to 7.
- Customer Type: Behavioral typology ("Phone Only", "Internet Only", "Bundle").
- Is Premium Customer: Binary indicator for Total Services $\geqslant 5$.
- Avg Monthly Charges: Ratio of Total Charges/tenure to compare between new and existing customers.
- Family Support: Indicator of family stability (Partner=Yes AND Dependents=Yes).
- Tenure Years: Conversion to years for improved business readability.

D. Predictive Modeling

1) Data preparation for learning: 17 categorical variables were transformed using label encoding. To prevent data leakage and ensure unbiased evaluation, we split the dataset into training (80%) and test (20%) sets before any transformation. The separation is stratified, ensuring the same proportion of churn in both sets (26,5%).

Standardization (centering and scaling) was applied to the numerical variables. An important step for algorithms sensitive to scale, although Random Forest is theoretically less sensitive in this regard.

2) Choice of algorithm: The choice of the random forest classifier is based on several theoretical and practical considerations, confirmed by recent literature. The authors of [9] showed that random forest and XGBoost offer better performance with 96% accuracy, while SVM and KNN offer lower performance. Also, in [10], random forest demonstrates higher performance on e-commerce data. Among the advantages of random forest are the following [11]:

- Robustness: Aggregating multiple trees reduces sensitivity to noise and outliers.
- Non-linearity: The model allows for the capture of complex relationships without relying on parametric assumptions.
- Interpretability: Provides a measure of variable importance.
- Stability: The average of sets produces a lower variance than that of single decision trees.
- Scalability: Good performance on moderately sized datasets.

3) Hyperparameter optimization: To identify the optimal configuration, an exhaustive search for hyperparameters was performed using gridsearchcv [12], with cross-validation (5 folds) and optimization based on the roc-auc metric:

- Number of estimators: [100, 200, 300]
- Maximum depth: [10, 20, 30, None]
- Minimum number of samples per leaf: [1, 2, 4]
- Split quality criterion: ['gini', 'entropy']

4) Evaluation metrics: To evaluate the model's performance, we used a set of complementary metrics:

- Accuracy: Gives the overall proportion of correct classifications.
- Precision: Measures the number of churners predicted who actually churn.
- Recall: Proportion of churn detected among all actual churn.
- F1-Score: Provides a balanced view by combining precision and recall.
- ROC-AUC: Model's ability to discriminate between classes.
- Confusion matrix: Offers visualization of prediction errors.

E. Deployment and Decision-Making Integration

1) Deployment architecture: After optimization, the final model was saved with all necessary components, e.g., scaler, encoders, and feature list, to ensure reproducibility for production deployment.

2) Streamlit application: For deployment, we propose using Streamlit, a Python framework that transforms scripts into interactive web applications without needing certain front-end skills. This step aims to make the models accessible to business teams [13].

The interactive web application offers:

- A form for inputting customer characteristics.
- Real-time calculation of churn probability.
- Risk level classification (Low, Medium, High) based on thresholds.
- Automated recommendations personalized to the risk profile.

3) Business intelligence integration: The goal of this phase is to structure the data and model predictions in a MySQL database and then connect them to Power BI. This architecture

combines descriptive analytics (BI) and predictive analytics (ML) within the same decision-making environment.

IV. Results and Analysis

A. Exploratory Analysis Results

1) Data structure and quality: The loading of the dataset exposed a structure of 7,043 rows and 21 columns, which is manageable for iterative processing. Analysis of variable types identified 18 categorical variables (object) and 3 numerical variables.

A variable "Total Charges" represents quantitative financial data, and we observed that it was stored in object format. This indicated that a conversion would be necessary.

2) Distribution of the target variable: Analysis of the churn variable indicates an imbalanced class configuration (see Fig. 2):

- Non-churn customers: 5,174 (73,5%)
- Churn customers: 1,869 (26,5%)
- Overall churn rate: 26,5%
- Class ratio: 2,8 (non-churn) to 1 (churn)

Moderate imbalances require careful attention during modeling, especially when selecting evaluation metrics.

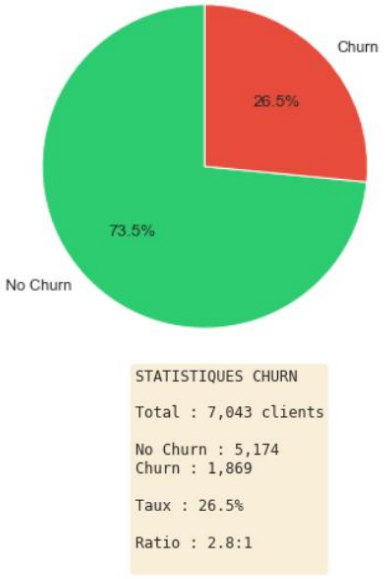


Fig. 2. Churn variable Analysis.

3) Discriminant factors of churn: Analysis identified a number of variables that were strongly associated with churn: (See Fig. 3, Fig. 4, Fig. 5, and Fig. 6).

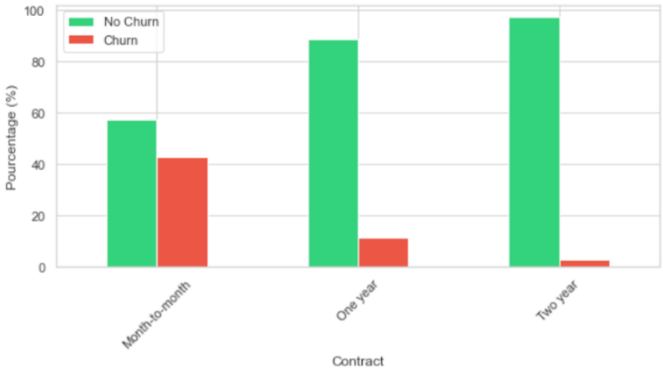


Fig. 3. Churn rate by contract.

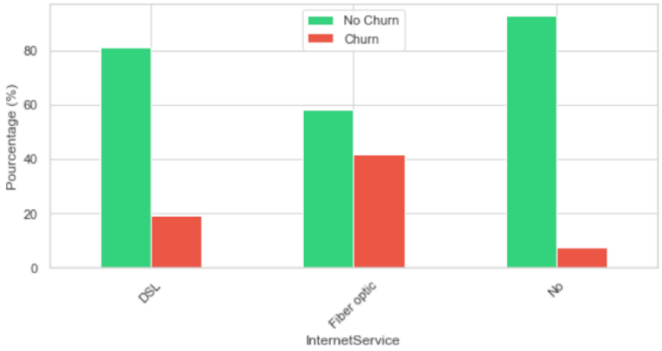


Fig. 4. Churn rate by internet service.

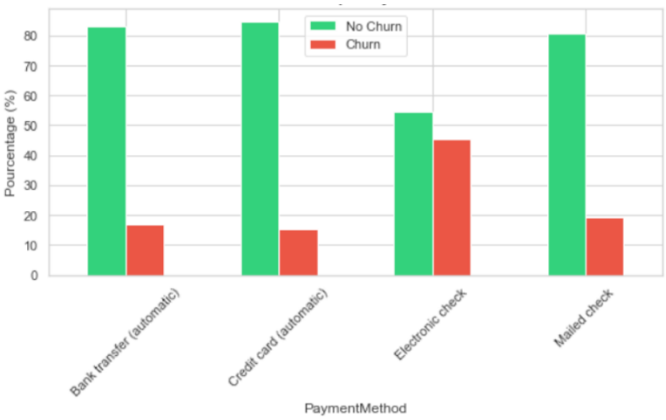


Fig. 5. Churn rate by payment method.

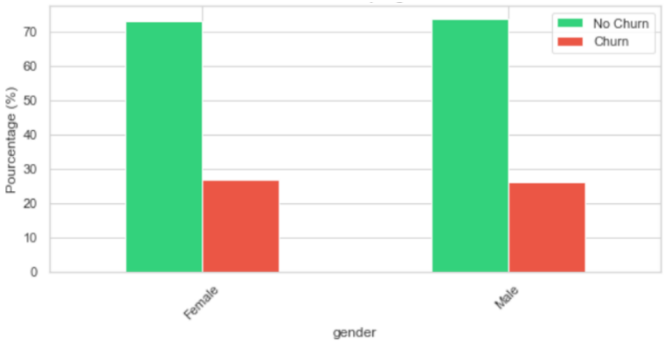


Fig. 6. Churn rate by gender.

a) Contract type: The number of customers who canceled their contracts was 42,7% for monthly customers, whereas those with one year and two years contracts have rates of only 11,3% and 2,8%. This difference highlights the effect of long-term commitments.

b) Internet service: With a churn rate of 41,9%, fiber optic customers depart more frequently than DSL users 19% and those without internet connection. This may reflect higher expectations among fiber users or more intense competition in this segment.

c) Payment method: Customers who pay by electronic check are highest chance of unsubscribing, with a rate of 45,3%. This rate is too higher than those paying by bank transfer at 26,5%, credit card at 21,9%, or by mailed check at 21%.

d) Gender: The difference is minimal, 26,1% of men churn and 26,9% of women, meaning that gender does not play a significant role.

4) Analysis of numerical variables: This analysis of continuous variables identified different profiles between churned and non-churned customers:

a) Tenure: Median duration for churned customers is 10 months compared to 38 months for non-churned customers.

b) Total Charges: Average among churned customers (1531$) is lower than the non-churned customers (2555$).

The correlation matrix (see Fig. 7) indicates a negative correlation between tenure and churn (-0,352), a positive correlation with Monthly Charges (0,193), and a weak correlation with Senior Citizen (0,151).

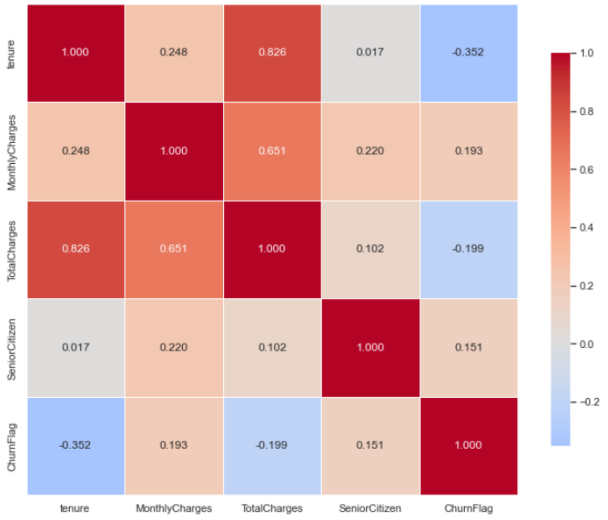


Fig. 7. Correlation matrix.

B. Modeling Results

1) Baseline model performance: The baseline random forest model, without optimization, attained the following performance on the test set:

- Accuracy: 79%
- Precision: 63%
- Recall: 48%
- F1-Score: 55%

These results provide a functional performance, but recall at 48% indicates a limitation. In other words, more than half of churned customers are not detected.

2) Optimized model performance: In order to improve the model performance, an optimization using gridsearchcv has been applied (see Table III):

TABLE III. PERFORMANCE COMPARISON BEFORE/AFTER OPTIMIZATION

Metric	Initial model	Optimized model	Improvement
Accuracy	79%	80%	+1,3%
Precision	63%	65%	+3,2%
Recall	48%	49%	+2,1%
F1-Score	55%	56%	+1,8%
ROC-AUC	0,83	0,85	+2,4%

Although these gains may seem modest, each additional percentage point of recall represents dozens of identified customers. Therefore, the improvement can be considered significant.

3) Confusion matrix analysis: The confusion matrix of the optimized model is illustrated in Fig. 8:

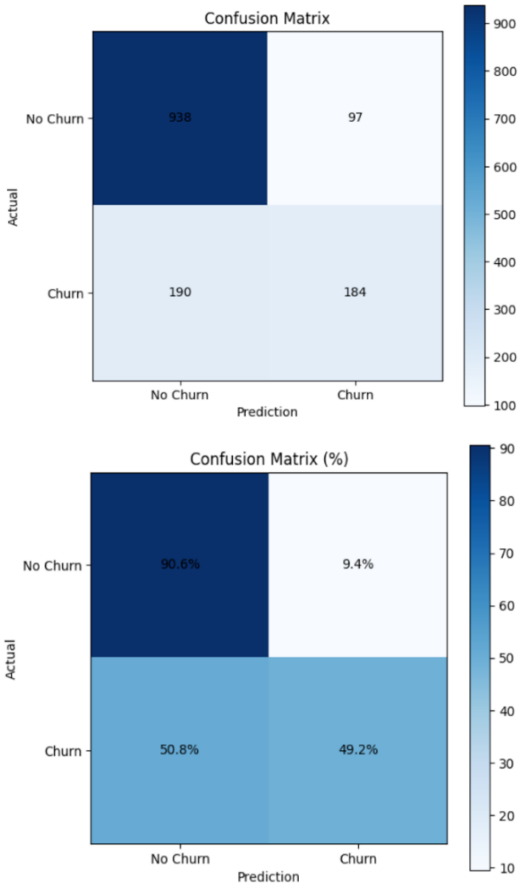


Fig. 8. Confusion matrix

The model prioritizes reducing false positives (97 loyal customers identified) at the expense of a higher recall rate (190 false negatives).

In an operational context, mobilizing retention resources to customers who will not leave (false positives) means spending without return, while failing to identify those who will leave (false negatives) means lost revenue.

4) ROC curve: The ROC curve presents an area under the curve greater than 0.84 (see Fig. 9). This indicates an excellent distinction capability of the model, and the value, above 0,8, is generally considered acceptable in real contexts.

The deviation from the random diagonal confirms that the model provides real added value compared to a random prediction.

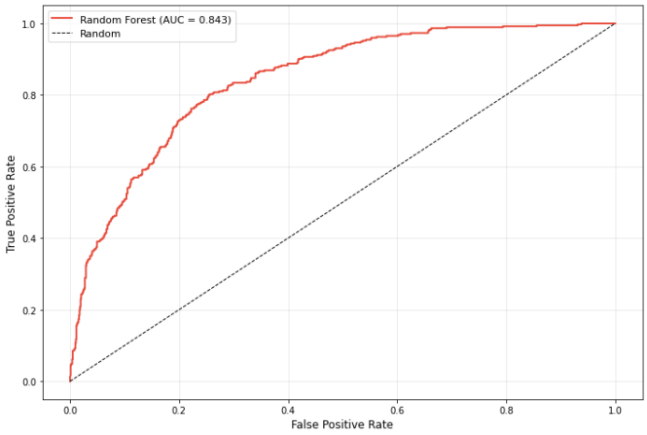


Fig. 9. Courbe ROC

5) Features importance: Fig. 10 illustrates the hierarchy of the most influential predictors detected during the analysis of variable importance.

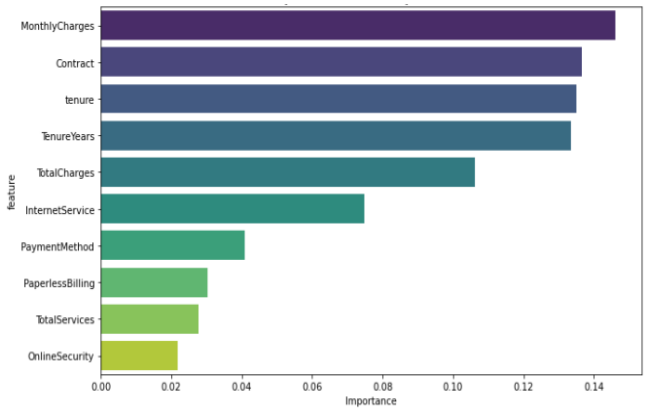


Fig. 10. Features importantes.

C. Operational Deployment Results

1) Streamlit application: Real-time interaction with the model is made possible through the application that we developed and usability tests have demonstrated:

- Less than 2 minutes on average for each customer's input.
- Intuitive comprehension of the results, such as probability and risk level.

The application was structured to support two different ways of using:

- Individual prediction for a specific customer (customer service).
- Batch prediction via file upload (marketing campaigns).

2) Power BI dashboards: Two dashboards were created to meet various decision-making needs (see Fig. 11, Fig. 12):

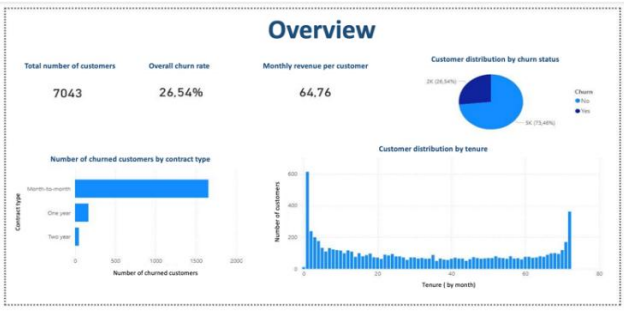


Fig. 11. Dashboard overview.

a) Overview: Designed for strategic management. It presents key performance indicators such as the number of customers, churn rate, average revenue, etc., and the classification of churn by contract and tenure. It provides an immediate overview of the portfolio.

b) Detailed analysis: Focused on marketing analysis. This report explores the relationships between churn and explanatory variables such as internet service, payment method, and cross-analysis of contract and payment. Churn evolution by tenure is also visualized. And a list of the top ten customers with the level of risk they pose.

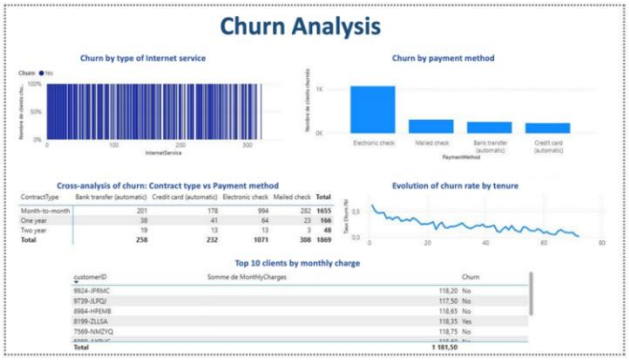


Fig. 12. Dashboard: Detailed Analysis.

V. Discussion

A. Interpretation of Results

The results obtained confirm and enrich the existing literature on churn prediction, while providing important nuances regarding the hierarchy of predictive factors.

Contrary to several studies that place tenure as the top predictor [4], our analysis reveals that MonthlyCharges constitute the most influential factor. This result suggests that price sensitivity is a major determinant of the cancellation decision.

Contract type ranks second, confirming the protective effect of long-term commitments. Month-to-month contract customers have a churn rate of 42.7%, compared to only 11.3% for one-year contracts and 2.8% for two-year contracts.

Tenure as a protective factor, but not preeminent. Although tenure ranks third, it does not dominate the hierarchy. This result nuances the idea of an automatic "lock-in" effect linked to relationship duration.

The presence of TotalServices in the top nine validates the relevance of the feature engineering work. The more services a customer consumes, the more engaged they are and the less likely they are to cancel.

B. Impact of Class Imbalance

The dataset exhibits a moderate class imbalance with a ratio of 2.8:1. While this challenge is identified in the literature [4], we chose not to apply oversampling or under-sampling techniques. Random Forest is inherently robust to moderate imbalance due to its bootstrap sampling and tree aggregation mechanisms. Furthermore, applying balancing techniques would alter the real-world class distribution, potentially leading to overoptimistic estimates that do not reflect operational conditions. The ROC-AUC of 0.85 is robust to imbalance, whereas the accuracy of 80% is partially inflated by the majority class.

C. Methodological Contributions

Our main methodological contribution is the integration of the data value chain from extraction to deployment. The feature engineering phase was particularly productive. A number of variables were created, such as IsPremiumCustomer and TotalServices, which emerged as important predictors of the final model. This confirms the intuition that the creation of relevant variables, guided by business knowledge, can be more decisive than the choice of algorithm.

The dual-channel deployment approach (interactive application for operational staff, dashboards for decision-makers) addresses a real need of organizations: making predictions available to different stakeholders according to their specific working methods.

D. Limitations and Perspectives

Our research, although having achieved its objectives, presents certain limitations that open up promising research directions.

1) Temporal limitation and longitudinal perspective: The study is based on a cross-sectional dataset, capturing churn at a given point in time. A temporal perspective would allow modeling the evolution of risk and identifying weak signals that precede churn. The use of longitudinal data could refine the understanding of the role of tenure and monthly charges, which occupy a central place in our variable hierarchy.

2) Informational limitation and enrichment perspective: The data used, although rich, are limited to the operator's internal information. The integration of external data like sentiment on social networks, macroeconomic indicators, and competitive data would constitute an important enrichment. Significantly, this addition makes it possible to understand why monthly charges become the primary predictive factor in our model by placing them within their competitive environment.

VI. Conclusion

This work aimed at designing and implementing a churn prediction system adapted to the telecommunications context. Rather than focusing only on modeling, we structured the approach around several successive steps, including exploratory data analysis, the development of an ETL process, feature engineering, and the use of a Random Forest model. The results were then integrated into visualization tools to support decision-making.

The obtained results show that the proposed approach is effective in identifying the main factors related to customer churn. In particular, variables such as monthly charges, contract type, and customer tenure appear to play a significant role. The feature engineering step also brought additional value. For instance, the variables TotalServices and IsPremiumCustomer, which were constructed during the process, turned out to be informative for the model.

One important point that emerges from this study is that predictive performance alone is not sufficient. In practice, the utility of such a model depends on how it is integrated into workflow, from data preparation to the interpretation of results. This perspective makes it possible to better align technical outputs with operational needs, and suggests that similar approaches could be applied in other contexts.

References

[1] F. F. Reichheld and W. E. Sasser, “Zero defections: Quality comes to services,” Harvard Business Review, vol. 68, no. 5, pp. 105–111, 1990.

[2] E. Ascarza, S. A. Neslin, O. Netzer, Z. Anderson, P. S. Fader, S. Gupta, et al., “In pursuit of enhanced customer retention management: Review, key issues, and future directions,” Customer Needs and Solutions, vol. 5, no. 1, pp. 65–81, 2018.

[3] S. Dhariya and H. Bhaidasna, “Ensemble learning approach for customer churn prediction in the telecom industry,” AIP Conference Proceedings, vol. 3288, no. 1, p. 040007, 2025.

[4] M. Imani and H. R. Arabnia, “Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning,” Machine Learning and Knowledge Extraction, vol. 7, no. 3, p. 105, 2025.

[5] Y. Koulou and N. El Hami, “Machine learning for recommender systems under implicit feedback and class imbalance,” International Journal of Advanced Computer Science and Applications, vol. 16, no. 9, 2025, doi: 10.14569/IJACSA.2025.0160954.

[6] D. Asif, M. S. Arif, and A. Mukheimer, “A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry,” Results in Engineering, vol. 26, p. 104629, 2025, doi: 10.1016/j.rineng.2025.104629.

[7] Y. Koulou, N. El Hami, A. El Attaoui, and S. Rhouas, "Application of Interoperability to Intelligent System of Systems," in Methods and Applications of Artificial Intelligence Dynamic Response Learning Random Forest Linear Regression Interoperability Additive Manufacturing and Mechatronics Volume 2. Open source preview, 2025, vol. 2, pp. 193-222.

[8] X. Su and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," Advances in Artificial Intelligence, vol. 2009, Art. no. 421425, 2009.

[9] B. Jagadeesh and K. Sashirekha, "Enhancing accuracy of customer churn prediction in e-commerce using random forest algorithm over K-nearest neighbors algorithm," AIP Conference Proceedings, vol. 3270, no. 1, p. 020101, 2025.

[10] M. A. Shaikhsurab and P. Magadum, "Enhancing customer churn prediction in telecommunications: An adaptive ensemble learning approach," arXiv preprint arXiv:2408.16284, 2024, doi: 10.48550/arXiv.2408.16284.

[11] N. Elsayed, S. A. Elaleem, and M. Marie, "Improving prediction accuracy using random forest algorithm," International Journal of Advanced Computer Science and Applications, vol. 15, no. 4, 2024, doi: 10.14569/IJACSA.2024.0150445.

[12] N. A. Alrajeh, B. Elmir, B. Bounabat, and N. E. Hami, "Interoperability optimization in healthcare collaboration networks," Biomedizinische Technik, vol. 57, no. 5, pp. 403-411, 2012.

[13] N. I. Yusof, N. M. M. Zainuddin, N. H. Hassan, N. N. A. Sjarif, S. Yaacob, and W. A. W. Hassan, "A guideline for decision-making on business intelligence and customer relationship management among clinics," International Journal of Advanced Computer Science and Applications, vol. 10, no. 8, 2019, doi: 10.14569/IJACSA.2019.0100865.

Attribute-Conditioned Attention Scaling for Text-to-Image Diffusion Models

Rabia Tahir, Bilal Ahmed Memon
School of Software Engineering, Tashkent University of Information Technology (TUIT), Tashkent, Uzbekistan
School of Business and Economics, Westminster International University, Tashkent, Uzbekistan

***Abstract*—Many text-to-image diffusion models have attained significant success in the generation of images from textual prompts; however, they still face some challenges, such as limited fine-grained control over different semantic attributes. To overcome this issue, this study proposes an Attribute-Conditioned Attention Scaling (ACAS), which modulates the cross-attention layers of the UNet model using attribute-related scaling factors. These scaling factors are assigned to the attention maps that allow selective enhancement of features in the ACAS without retraining. Moreover, this model allows precise control over generated images without retraining of the base model at the inference level, along with preservation. For experiments, 30 diverse prompts along with eight descriptive attributes are used to inspect the multi-attribute controllability of the proposed model. Different evaluation metrics, such as CLIP, LPIPS, and Inception Score (IS), are used to quantitatively evaluate the proposed model. Experimental results prove that the proposed model ACAS obtains competitive results with an LPIPS of 0.75, CLIP score of 0.316, IS of 3.98, and a minimal 8.47 seconds generation time. Furthermore, a comparative analysis of the ACAS model with similar baseline methods is performed, and the comparison shows that the ACAS improves attribute controllability without adding extra computational cost. Overall, this model bridges the gap between fine-grained attribute control and prompt-based guidance in the latest diffusion models.**

***Keywords*—Diffusion models; attribute-control; CLIP; LPIPS; text-to-image generation; cross-attention modulation**

I. Introduction

In today's world, numerous people want to create images inspired by their visual imagination. With the advent of text-to-image models, it is possible to generate visually stunning images by just providing a prompt [5]. Diffusion models have emerged as a powerful method in the domain of text-to-image generation, and they are inspired by non-equilibrium thermodynamics. The rising interest in diffusion models proves their versatility and potential in various image-related tasks, especially text-to-image generation [1]. Latest models such as DALLE-2 [2], Stable Diffusion [10], and Imagen [4] prove the power to generate photorealistic and diverse images by conditioned-based prompts. Despite the huge success of these models, there are still a few challenges, such as limited fine-grained control on semantics attributes, e.g., "younger face", "metallic texture", and "futuristic city". Existing models are aimed at two main directions, i.e., fine-tuning the model (retraining) and embedding-level guidance. The fine-tuning or retraining of the model leads to more computational cost, while embedding-level guidance lacks explicit mechanisms to adjust the attributes' influence during the denoising task. The main motive of the proposed model is to modulate the influence of attributes on the attention layers with the help of a scalar weight. This allows suppression, increase, and balance of the attribute strength in the image generation. Further, the proposed model, ACAS, can deal with multiple attributes simultaneously with the help of normalized weights. Hence, the proposed model can integrate diverse attributes like ("younger", "hat", "smiling") in one-step generation. To address these issues, this study proposes a lightweight Attribute-Conditioned Attention Scaling (ACAS) for multi-attribute-based text-to-image generation in a well-controllable manner with no retraining overhead. During the inference stage, the proposed model injects attribute-specific scaling coefficients into the cross-attention blocks of the UNet model. The main motive is to allow all attributes to be balanced and enhanced while generation with the help of a scalar control parameter. The proposed model offers three significant benefits: fine-grained control, multi-attribute conditioning and no retraining overhead. Experiments are performed and a comparison is conducted for the proposed model with baseline models and the results show the superiority of the proposed model. The results are evaluated with both quantitative and qualitative metrics, including CLIP [15], LPIPS [16], Inception Score (IS) [17] and generation time. Overall, the contributions of this study are:

- A novel ACAS model is proposed for controllable attribute-guided text-to-image generation without retraining.
- This model injects the semantic attributes into the denoising to preserve the textual prompt semantics and to enhance the fidelity of attributes.
- Extensive experiments are performed with a comprehensive evaluation to show the efficiency of the proposed model.
- The proposed model, ACAS, bridges the gap between fine-tuning and prompt-level control. It offers a flexible and practical solution for real-time text-to-image generation.

The rest of the study is organized as follows: Section II describes the related work on controllable text-to-image generation and diffusion-based methods. Section III explains the proposed Attribute-Conditioned Attention Scaling (ACAS) model. Section IV illustrates the data, experimental setup, quantitative and qualitative results, discussion, and ablation study. Finally, Section V concludes the study and highlights some future research directions.

II. Related Work

Generative Adversarial Networks (GANs) have been used for image generation and they achieved strong performance on the benchmark datasets. However, these models even suffer from training instability and limited controllability limitations [3]. Diffusion-based text-to-image methods [6, 7, 8, 10] have shown significant progress in the past few years by generating highly realistic images with the help of a text prompt. Diffusion-based models avoid the hectic process of training a huge text-to-image model from the beginning and take benefit of already existing pre-trained models for attribute manipulation and image editing tasks [21]. Despite their huge success, some models [4, 9, 23] lead to high computational cost, and many of these models depend on denoising diffusion probabilistic models (DDPM).

Vanilla stable [10] diffusion framework proposed by Romach et al., produces images from textual prompts with the help of a latent diffusion manner. Instead of generating high-quality images, this model has low control over fine-grained attributes like environment, style type, and mood. Moreover, Prompt-to-prompt [11] and ControlNet [12] show improved guidance for generated images, but they still need extra control with specific attribute adjustment. Uni-ControlNet [18], the extended version of ControlNet, allows various control models in a unified framework simultaneously and leads to composite local and global controls. Furthermore, DynamicControl [19] improves adaptive selection and ordering of multiple control signals in order to handle various conditions. It improves the fine-grained attribute adjustment in distinct text-to-image generation tasks. These models either need additional input or are less efficient to control various attributes simultaneously. DiffEdit [14] and Attend-and-Excite [13] models perform attribute-level manipulation as well, but they require complicated pipelines for unseen attributes. Latest research explores the semantic and fine-grained control in a more significant manner. DilightNet [20] proposed fine-grained lightning control with the help of diffusion process guidance. This method produces images with respect to a given text prompt and lightning. By keeping these limitations in mind, Attribute-Conditioned Attention Scaling (ACAS) is designed to control multiple attributes over generated images in a precise way without retraining the base model. To check the efficiency of the proposed model, ACAS, state-of-the-art evaluation metrics are used, i.e., CLIP score [15], LPIPS [16], Inception score (IS) [17], and generation time. These metrics are used to evaluate the perceptual diversity, semantic alignment and generated content quality. Besides, the comparison of the proposed model, ACAS, with related models is performed, and the results show the superiority of the model in various factors.

III. Methodology

A. Overview

The proposed model, Attribute-Conditioned Attention Scaling (ACAS), improves the traditional text-to-image diffusion methods in order to obtain precise control over visual attributes in the produced image without retraining. As compared to the vanilla diffusion model that globally interrupts the prompt, this model applies a forward hook added at all cross-attention layers of the UNet without retraining. A prompt P and a set of descriptive attributes $A = \{a_1, a_2, a_3, \ldots, a_n\}$ are assigned to ACAS as an input. Then, the proposed model modulates the UNet cross-attention map M during the forward pass, as shown in Eq. (1) below:

$$\tilde{M} = \gamma\,(a_i).M \tag{1}$$

where, $\gamma(a_i)$ is the attention scaling factor for attributes a_i, M is UNet based cross attention map and $\tilde{M}$ is the modulated attention map that is used for forward pass after scaling. $\gamma(a_i)$ allows controllability of attribute influence without the need of retraining the base model. Fig. 1 shows the systematic view of the proposed model ACAS.

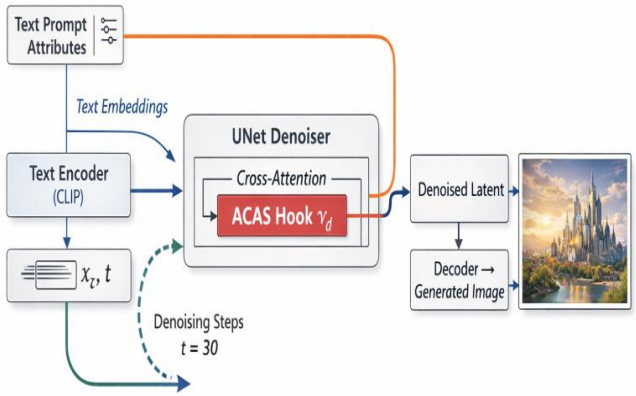


Fig. 1. Systematic view of the proposed model ACAS.

B. Cross-Attention in Traditional Diffusion Models

UNet in the traditional text-to-image diffusion models uses cross-attention to combine latent image features f_x with text features f_t, as explained in Eq. (2) below:

$$\text{Att(K, V, Q)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{2}$$

where, K, V and Q are key, value and query metrics that are derived from text and image immersing.

C. Scaling by Attributes

To modulate attention weights, the proposed model ACAS presents a scalar $\gamma(a_i)$, as shown below in Eq. (3):

$$\widetilde{Att}\text{(K, V, Q; } a_i) = \gamma(a_i).\,\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{3}$$

where, $\gamma(a_i) < 1$ subdues the features influence and $\gamma(a_i) > 1$ improves it. During denoising in UNet, it is imposed in each cross-attention block.

D. Multi-Attribute Scaling

In real-time, users give more than one attributes in a textual prompt to generate an image. For example, "an old man is smiling and wearing hat". Therefore, a single scalar is not enough to address this issue. The proposed model introduces a weighted multi-attribute formulation. Suppose, there is a set of N attributes $A = \{a_1, a_2, a_3, \ldots, a_n\}$, where a scalar weight is assigned to each attribute and this weight is independent. Hence, this multi-attribute attention is defined in Eq. (4):

$$\widetilde{Att}\text{(K, V, Q; } A) = \frac{1}{\sum_{i=1}^{N}\gamma(a_i).}\sum_{i=1}^{N}\gamma(a_i).\,\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

where, $\sum_{i=1}^{N} \gamma(a_i)$ is a normalization term that handles the magnitude of attention when more attributes are added. It allows the attributes to be scaled independently and balanced with weight selection.

E. Integration into UNet Denoising Process

Stable diffusion addresses the iterative refinement for a latent noise vector z_t during the denoising step. In the proposed ACAS, the modified attention $\widetilde{Att}(K, V, Q; A)$ is used at all cross-attention blocks of the base UNet model. This modifies attention and applies during all denoising step $t \in \{1, 2, \ldots, T\}$. Suppose, $\epsilon_\theta(z_t, t, P)$ shows a denoising UNet with condition oriented prompt P. Then, with ACAS, the conditioning prompt is shown in Eq. (5):

$$\epsilon_\theta(z_t, t, P, A) = \epsilon_\theta(z_t, t, P, \widetilde{Att}(K, V, Q; A)) \quad (5)$$

Afterward, the latent space becomes:

$$z_{t-1} = \text{Denoise}(z_t, \epsilon_\theta(z_t, t, P, A)) \quad (6)$$

This step in Eq. (6) has no training overhead and it does not change the architecture of UNet.

F. Weight Interpretation and Optimization

The attribute weightage $\gamma(a_i)$ gives a balanced control to influence the attributes. $\gamma(a_i)$ controls the degree of attention given to the attribute by the model in the cross attention map. Higher value of $\gamma(a_i)$ means a stronger influence, while low value of $\gamma(a_i)$ leads to lower influence. This degree of control over multiple attributes is without retraining the model. It is described below:

$\gamma(a_i)<1$ decrease the influence

$\gamma(a_i)>1$ increase the influence

$\gamma(a_i)=1$ same like vanilla diffusion model

Moreover, the proposed ACAS allows learnable weights for strong semantic alignment. $\gamma(a_i)$ is optimized with a CLIP-based loss, as shown in Eq. (7):

$$L_{CLIP} = 1\text{-}COS(f_T(P, A), f_I(A)) \quad (7)$$

where, $f_T(.)$ and $f_I(.)$ are CLIP-based text and image encoders and COS is used for cosine similarity. This loss function helps to generate an image which is aligned with both the attributes and prompt.

IV. Experiments

A. Data

For evaluation, 30 diverse prompts with eight attributes are used to generate 240 images. Fig. 2 shows some of these generated images with their prompt and attribute. For experiments, the prompt-based image generation is used as the base dataset. A total of 30 base prompts are defined that cover various visual domains such as scenes, human figures, animals and objects. For example; these textual prompts are "a futuristic city at sunset", "a surreal dreamscape". These prompts can be extended to different semantic categories that show the model's ability to create the diverse content with the help of multiple attribute as well. Against each prompt, eight semantic or stylistic attributes are used i.e. "low details", "high details", "mysterious", "cartoon", "hot and sunny", "raining", "realistic" and "oil painting". A strength coefficient (c) is assigned to each attribute that scales it in the attention layer of the UNet model while image generation. Table I shows the different value of weightage attribute γ that are assigned to all attributes. For example, $\gamma > 1$ is used for those attributes that require higher stylistic features, such as cartoon and high detail. For $\gamma < 1$, it implies that attribute details are not dominating the text prompt, such as "Low detail". The γ values are selected regarding the expected visual dominance of each attribute. It is ideal to start the initial estimate like 0.5 and then increase it or decrease it according to the obtained results in order to get the balance fine-grained control of attribute influence. This strength value gives fine-grained control over attributes. Table I shows the best γ values that yielded best in the experiments.

The dataset used for experiments in this work is designed to evaluate attribute-level controllability instead of large-scale image generation. Each attribute-prompt pair is constructed carefully in order to cover diverse semantic categories. These pairs cover diverse semantic categories, which are style, appearance and object-level characteristics. The applicability of the ACAS is not limited to a specific dataset because it works by modifying the cross-attention layer within a diffusion model. If the number of prompts increases, it would not alter the final results significantly. The consistent performance against diverse attribute-prompt pairs shows that the model ACAS generalizes across multiple visual concepts.

TABLE I. Strength Values Used in the Model Against Each Attribute.

Attributes	Strength (γ)
Low Detail	0.75
Realistic	1.25
High Detail	1.35
Cartoon	2
Hot and Sunny	1.15
Rainy	1.50
Mysterious	1.10
Oil Painting	1.05

B. Model Setup

For baseline method, Stable Diffusion v1-5 from Runway ML is used for image generation. Experiments are performed on Google Colab with T4 GPU and mixed precision (torch.float16) to optimize inference speed and memory. Mixed-precision helps in reducing memory without degrading the semantic quality. For comparison, all images are generated with fixed diffusion steps, i.e., 30 for ACAS, baseline and vanilla diffusion model. Moreover, CLIP (ViT-B/32) is used for quantitative evaluation of semantic alignment among the textual prompts and generated content. Table II shows details of parameters used in the proposed model, ACAS.

Contrastive Language Image Pre-training (CLIP) is a model proposed by OpenAI that is used to compare the images and text in the likewise semantic space. It calculates how the image matches with a text prompt. Vit-B/32 is a version of CLIP that

uses Vision Transformer as a backbone in the model. It deals with the images in patches of 32x32 pixels. So, this is used as an automatic metric to calculate the score for how well the output images matches the text description. A forward hook is implemented on all cross-attention layers of UNet model in order to modulate the influence of attributes on the generated image. With the help of attribute strength factor γ, the scaling of attention outputs can be controlled. Each text prompt is concatenated with the attribute, such as "a retro television set + high detail". After concatenation, pass it to the stable diffusion pipeline. The generation time is calculated for each image and stores the image in the corresponding directory.

TABLE II. PARAMETERS AND THEIR VALUES USED IN ACAS

Parameter	Value / Setting
Base model	stable-diffusion-v1-5
Diffusion Steps	30
Image Size	512 × 512
Device	Cuda (GPU)
Attention Hook	Custom Cross Attention hook
Mixed Precision	torch.autocast("cuda")
CLIP Model	ViT-B/32
LPIPS Model	Alex
Number of Prompts	30

C. Qualitative Results

This section shows the qualitative results of the proposed model. Fig. 2 shows the generated images by different textual prompts with eight different attributes. These attributes are “cartoon”, “high details”, “low details”, “mysterious”, “oil painting”, “photorealistic”, “Rainy” and “hot and sunny”. This figure shows results against four prompts which are “are "a futuristic city at sunset", “a tropical beach with palm trees", "a medieval castle on a hill" and “a cup of coffee”. In Fig. 3, the proposed model, ACAS, generates appealing images that describes both text prompt and attributes. For example, “a futuristic city at sunset” with cartoon attribute reflects an image that represents the futuristic city scape with cartoon aesthetic. This demonstrates the capability of ACAS to disentangle and recombine the attribute and textual information. Fig. 3 shows the visual comparison of the proposed model with the baseline stable diffusion v1.5 model [10] and the vanilla diffusion model v1.4 [10]. This clearly depicts the superiority of the proposed model over baseline models. The first row shows "a medieval castle on a hill" with the attribute “Oil painting”. The ACAS model generates a better image that satisfies both the text prompt and the attribute. The second row shows "a white dove flying” with the attribute “rain”. Vanilla and base model failed to generate a realistic image with a complete dove flying in rain, but ACAS generates an image that portrays the actual text prompt and attribute. The last row shows "a robot holding a flower" with attribute “hot and sunny”. The other two models do not satisfy the “hot and sunny” attribute and create an ordinary image of robot holding a flower. However, the proposed model generates a realistic image of a robot holding a flower in hand while the background shows the hot and sunny weather fulfilling the attribute condition. This is possible because the proposed model efficiently combines the attribute-guided attention with the text prompts and precisely controls the visual features of the generated image. This makes it a promising model for generating stylist images that reflects both textual prompts and attributes.

D. Quantitative Results

This section demonstrates the quantitative results for the proposed model, ACAS, and with related models stable diffusion v1.5 [10] and vanilla diffusion v1.4 [22]. State-of-the-art evaluation metrics CLIP [15], LPIPS [16], Inception score (IS) [17] are used along with the generated time for each image. A total of 30 experiments are performed for each model to generate 30 different images with respect to the attribute and text prompt and an average is calculated for all experiments. Table III summarizes these results for all three models. To obtain fair results, the same parameter values are used for all three models. The CLIP score is used to calculate the semantic consistency among generated images and text prompts. To calculate this score, OpenAI CLIP ViT-B/32 is used for calculating cosine similarity between image features and the encoded text. The second measure is LPIPS, which calculates the perceptual diversity among generated images with the help of an AlexNet Backbone. Pairwise LPIPS distances are measured for each attribute and then averaged. The Inception Score (IS) is used to inspect the quality and diversity of the generated images. The higher value of IS shows better results and is consistent to the textual prompt and attribute. The results in Table III show that ACAS obtained comparable results as compared to Stable Diffusion v.15 and vanilla diffusion v1.4. These results show that the proposed model, ACAS, preserves the overall quality of images without degradation. Furthermore, it is noted that these standard metrics are designed to compute the perceptual similarity, global image quality and distribution of realism. For fine-grained attribute control, they are not sensitive. Thereby, the lack of significant improvements in these standard metrics is not surprising and does not contradict the novelty and contribution of the model. The main motive of ACAS lies in allowing interpretable attribute-specific control by using attention scaling in the cross-attention layers instead of optimizing these standard metrics. Therefore, introducing controllable image generation without retraining with comparable performance to the strong baseline methods is a significant contribution. The generation time remains same for all models means that utilizing of attention hook in ACAS. Fig. 4 shows that the evolution of different metrics over time with respect to three models does not increase the computational cost additionally.



Fig. 2. This figure shows four different prompts with eight attributes. These prompts are "a futuristic city at sunset", "a tropical beach with palm trees", "a medieval castle on a hill" and "a cup of coffee". Attributes are given at the top of each column.

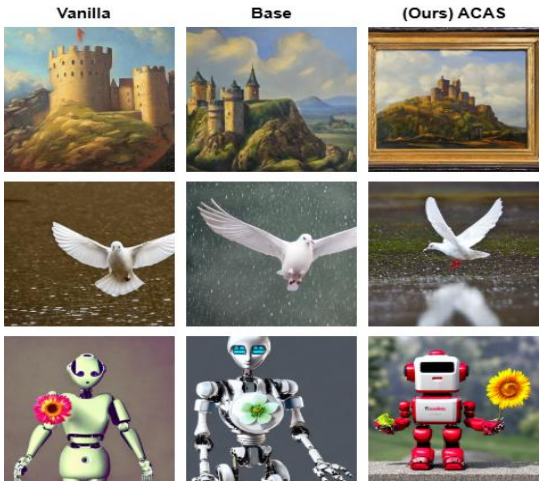


Fig. 3. Comparison of the ACAS with baseline models; first row shows "a medieval castle on a hill" + Oil painting, second row shows "a white dove flying" +rain, and the last row shows "a robot holding a flower" + hot and sunny.

TABLE III. SUMMARY OF CLIP, IS, LPIPS AND GENERATION TIME

Metrics	Vanilla V1.4	Stable Diffusion v1.5	ACAS (the proposed model)
CLIP	0.315	0.319	0.316
LPIPS	0.74	0.76	0.75
IS	3.94	4.06	3.98
Generated Time	8.4 sec	8.47 sec	8.47 sec

E. User Survey

A user survey is conducted to inspect the semantic alignment of the proposed model. Fig. 5 shows the summary of user survey to evaluate the performance of three models: ACAS, base model and vanilla diffusion. Five different textual prompts integrated with five attributes were used to generate images with all three models. Then, users were asked to choose the best image that matches both prompt and attribute. It is clearly evident in Fig. 5, that ACAS consistently obtained the highest score as compared to the other two models. The vanilla and base models were less selected and received fewer votes. This survey gives an overall insight and significance of the proposed model, ACAS.

F. Discussion and Ablation Study

The proposed model, ACAS, performance is discussed in this section. Three attribute-prompt pairs are used to inspect the effect of attribute strength gamma. These pairs are "a horse running in meadow", "a busy street with people" and "a waterfall in a dense forest", with three attributes "rainy", "mysterious" and "cartoon". Three different γ values, i.e., 1, 2 and 3 are used for the attribute strength.

The results show that when γ =1, the influence of the attribute is subtle and it preserves the main identity of the prompt. When γ increase to 2, then the visibility of attribute is more visible. When γ =3, then the attribute becomes more dominant and its influence is more visible on the generated images. However, increasing γ value more than 3 may lead to generate unrealistic images. As shown in Fig. 5, when γ is 3 then there are artifacts on the horse image in contrast to gamma value 2 that generates a balanced images. This observation shows that ACAS allows gradual and interpretable control over the attribute strength. In contrast to baseline methods that modify latent representation or need fine-tuning, the proposed model ACAS maintains the image quality while preserving the original semantics.

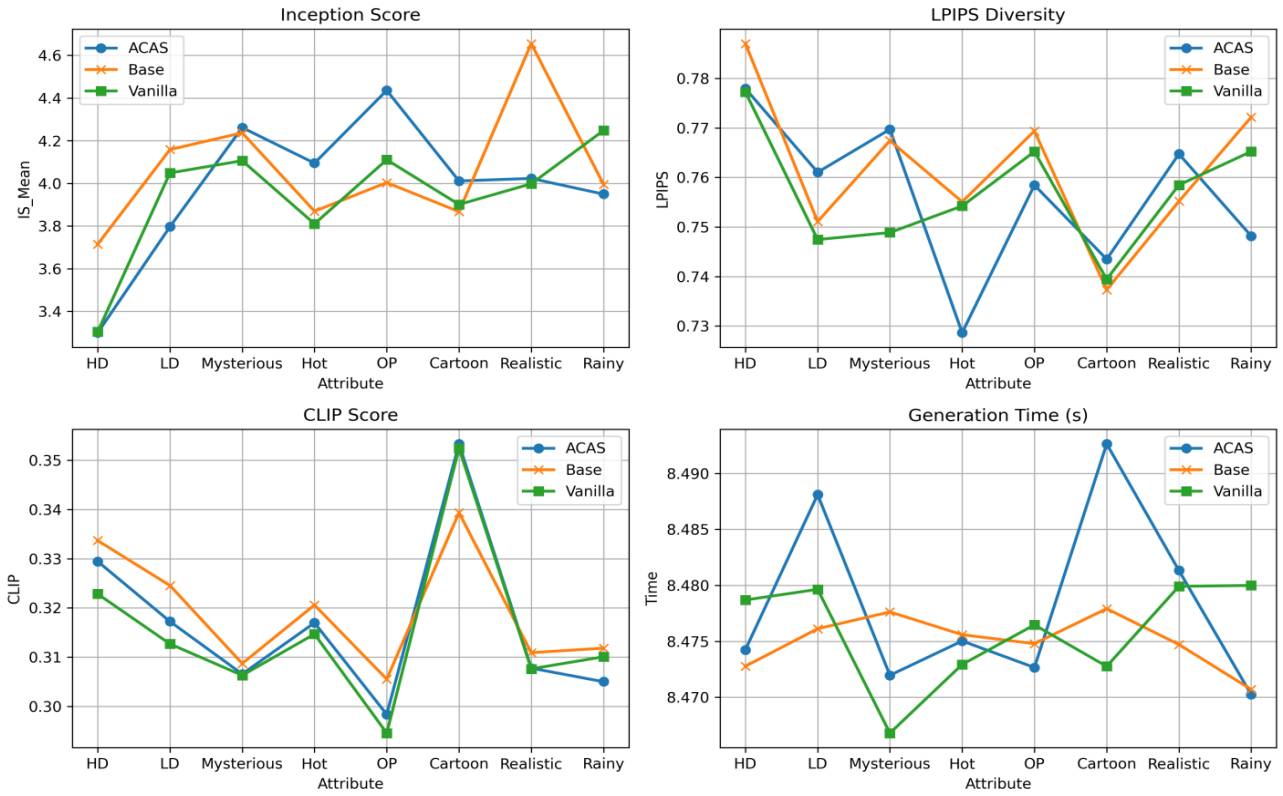


Fig. 4. Evolution of CLIP, LPIPS, and IS over time, along with generation time for each image (ACAS: the proposed model, Base: Stable diffusion v1-5, Vanilla).

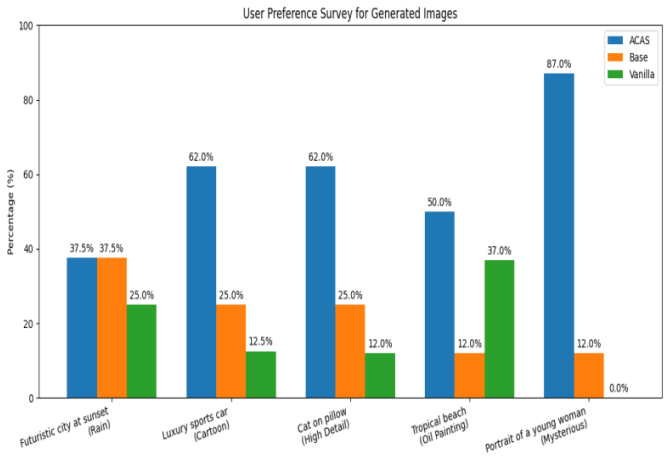


Fig. 5. User survey for ACAS, base model and vanilla model on generated images.

Moreover, this analysis shows that the standard quantities metric values are comparable with baseline models because the global image distribution is mainly unchanged during adjustment of attribute-specific features. In a nutshell, the ablation study shows that the ACAS provides interpretable and stable control over attributes. Furthermore, this discussion shows that the proposed model ACAS obtains controllability over attribute along with comparable performance to the similar existing models. In this section, the significance of attribute strength γ in the proposed model, ACAS, is briefly discussed. For this purpose, an ablation study is performed, focusing on attribute strength γ on the resultant images.



Fig. 6. Ablation study with different attribute strength γ ={1, 2, 3}. The first column shows generated images with $\gamma = 1$, the second column shows $\gamma = 2$ and the last column shows $\gamma = 3$.

Fig. 6 shows three different textual prompts with three different attributes. These prompts are "a horse running meadow", "a waterfall in a dense forest" and "a busy street with people" while the attributes are "rainy", "cartoon" and "mysterious". For each prompt-attribute pair, the attributes strength values are set to 1, 2 and 3. The generated images are obtained by just changing these values without any extra modification in the model. Fig. 6 shows that when γ =1, the effect of that specific attribute is subtle and not very dominant. When γ increases to 2 then the attribute influence becomes more

visible and prominent. At γ =3, the generated image strongly influences the textual prompt by intensified attribute details such as rain or cartoon textures. However, a higher value of γ does not guarantee the best results always. For example, when γ =3 then the image generated against the prompt "a waterfall in a dense forest" looks unrealistic and best results can be seen at γ =2. Therefore, it is concluded that ACAS allows interpretable control on various attributes without modifying the baseline model.

V. Conclusion and Future Work

This study presented a lightweight Attribute-Conditioned Attention Scaling (ACAS) to generate images with textual prompts and attributes. The selective enhancement of the features is made possible by applying the cross-attention layers of the UNet model along with attribute scaling factors. This enhancement of selective features against each attribute is achieved without retraining the base model. A total of four quantitative metrics, semantic alignment (CLIP), LPIPS, Inception Score (IS) and generation time are used. Results show that the descriptive attributes play a great role in generating behavior. Overall, the proposed model is highly responsive to attribute-based prompts and realism. In a nutshell, ACAS bridges the gap between textual prompt guidance and fine-grained attributes and presents a flexible model for controllable image generation with diffusion models. In the future, the aim is to extend this work by handling more complex and multiple attribute conditions and high-resolution image generation. Moreover, larger user studies and the use of more robust metrics could validate and enhance the performance of the model.

Acknowledgment

The authors would like to acknowledge the Tashkent University of Information Technologies (TUIT) and the Westminster International University in Tashkent, Uzbekistan for providing academic support and research facilities. Their institutional resources and environment contributed to the successful completion of this work.

References

[1] Y. Huang, J. Huang, Y. Liu, M. Yan, J. Lv, J. Liu, and S. Chen, "Diffusion model-based image editing: A survey," IEEE Trans. Pattern Anal. Mach. Intell., 2025.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," arXiv preprint arXiv:2204.06125, 2022.

[3] K. Cheng, R. Tahir, M. Li, and Z. Wang, "An analysis of generative adversarial networks and variants for image synthesis on MNIST dataset," Multimedia Tools and Applications, vol. 79, no. 19, pp. 13725–13752, 2020.

[4] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," Adv. Neural Inf. Process. Syst., vol. 35, pp. 36479–36494, 2022.

[5] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 3836–3847, 2023.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Adv. Neural Inf. Process. Syst., vol. 33, pp. 6840–6851, 2020.

[7] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," Adv. Neural Inf. Process. Syst., vol. 34, pp. 8780–8794, 2021.

[8] Y. Li, H. Wang, Q. Jin, J. Hu, P. Chemerys, Y. Fu, and J. Ren, "SnapFusion: Text-to-image diffusion model on mobile devices within two seconds," Adv. Neural Inf. Process. Syst., vol. 36, pp. 20662–20678, 2023.

[9] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, and M. Liu, "eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers," arXiv preprint arXiv:2211.01324, 2022.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 10684–10695, 2022.

[11] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross-attention control," arXiv preprint arXiv:2208.01626, 2022.

[12] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 3836–3847, 2023.

[13] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-Excite: Attention-based semantic guidance for text-to-image diffusion models," ACM Trans. Graph., 2023.

[14] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "DiffEdit: Diffusion-based semantic image editing with mask guidance," arXiv preprint arXiv:2210.11427, 2022.

[15] A. Radford et al., "Learning transferable visual models from natural language supervision," arXiv preprint arXiv:2103.00020, 2021.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 586–595, 2018.

[17] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training generative adversarial networks," Adv. Neural Inf. Process. Syst., vol. 29, pp. 2234–2242, 2016.

[18] S. Zhao et al., "Uni-ControlNet: All-in-one control to text-to-image diffusion models,"Adv. Neural Inf. Process. Syst., vol. 36, pp. 11127–11150, 2023.

[19] Q. He et al., "DynamicControl: Adaptive condition selection for improved text-to-image generation," arXiv preprint arXiv:2412.03255, 2024.

[20] C. Zeng et al., "DiLightNet: Fine-grained lighting control for diffusion-based image generation," ACM SIGGRAPH Conf. Papers, pp. 1–12, 2024.

[21] Z. Zhang et al., "SINE: Single image editing with text-to-image diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6027–6037, 2023.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proc. CVPR, 2022.

[23] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 4195–4205, 2023.

Adaptive Multi-Layer Encryption for Enhancing Security in Smart Home IoT Ecosystems

Dancan Obuya Machuki[1], Kennedy Ronoh[2]
Research Scholar, School of Computing and Engineering, Strathmore University, Nairobi, Kenya[1]
Senior Lecturer, School of Computing and Engineering Sciences, Strathmore University, Nairobi, Kenya[2]

***Abstract*—The proliferation of Internet of Things (IoT) devices in smart home environments has introduced significant security challenges, largely due to device heterogeneity, constrained computational resources, and limited native support for robust encryption mechanisms. This study presents an Adaptive Multi-Layer Encryption (AMLE) model designed to enhance security across the device, network, and cloud layers of smart home IoT ecosystems. The proposed model integrates lightweight encryption mechanisms at the device layer, machine learning-based anomaly detection at the network layer, and strong cryptographic protection combined with Attribute-Based Access Control (ABAC) at the cloud layer. Evaluation was conducted in a controlled, simulated environment to assess functional correctness, security behavior under representative threat scenarios, and system performance. Results demonstrate that the AMLE framework is capable of detecting and responding to simulated unauthorized access attempts, anomalous traffic patterns associated with botnet-like behavior, and data exfiltration scenarios, while maintaining operational performance suitable for typical smart home use cases. The Isolation Forest algorithm, configured with a contamination threshold of 0.05, successfully identified deviations from baseline traffic behavior and triggered policy-driven security responses. The findings indicate that AMLE provides a practical reference framework for implementing adaptive, layered security controls in smart home IoT environments, balancing security requirements with operational constraints.**

***Keywords*—Smart home security; IoT encryption; adaptive security; multi-layer protection; resource-constrained devices**

I. Introduction

Smart home IoT ecosystems consist of interconnected devices such as sensors, cameras, thermostats, and smart locks that collectively automate and enhance residential environments. While these systems offer increased convenience and functionality, their rapid adoption has intensified security challenges related to confidentiality, integrity, and access control [1]. A defining characteristic of smart home IoT devices is their constrained computational and memory resources, which limit the feasibility of implementing conventional cryptographic mechanisms commonly used in traditional computing environments [2].

As a result, many IoT devices deploy weak or inconsistent encryption, exposing them to threats including unauthorized access, man-in-the-middle (MitM) attacks, and botnet recruitment [3]. The heterogeneity of smart home devices further complicates security implementation, as devices vary widely in processing capabilities, communication protocols, and data sensitivity [4]. Prior studies identify insufficient encryption and weak access control as primary contributors to data exposure and privacy breaches in IoT environments [5].

Unencrypted or poorly protected communication between devices and cloud services enables attackers to intercept sensitive data, manipulate device behavior, or exploit compromised nodes to launch distributed attacks [6]. Botnets leveraging insecure IoT devices have been shown to significantly disrupt network reliability and compromise user safety [7].

A. Limitations of Existing Solutions

Existing cryptographic solutions face limitations when applied to resource-constrained IoT contexts. AES, although efficient in general-purpose systems, may impose computational overhead and complex key management requirements on low-power devices [8]. RSA supports secure key exchange but introduces high computational costs and latency unsuitable for frequent IoT communications [9]. Elliptic Curve Cryptography (ECC) reduces key sizes while maintaining security strength but presents challenges related to computational complexity, curve selection, side-channel vulnerabilities, and key management [10].

Beyond cryptography, traditional network security mechanisms such as firewalls and signature-based intrusion detection systems often lack adaptability and granular control suitable for IoT environments [11]. Signature-based approaches are particularly ineffective against previously unseen attacks, limiting their applicability in dynamic and evolving threat landscapes [12].

B. Research Contribution

This research addresses these challenges by proposing an Adaptive Multi-Layer Encryption (AMLE) model tailored for smart home IoT ecosystems. The model applies layered security controls across the device, network, and cloud layers, with encryption strength and security responses adjusted based on predefined policies that consider device capabilities, data sensitivity, and detected threat indicators.

The remainder of this study is organized as follows. Section II reviews related work on IoT security mechanisms and identifies the gap this study addresses. Section III describes the research methodology and system architecture. Section IV presents the evaluation results. Section V discusses the findings, and Section VI concludes the study.

II. Related Work

Research on smart home IoT security has explored encryption, anomaly detection, and access control as distinct mechanisms, yet integrated frameworks that adapt security

posture dynamically across all layers remain limited. The studies reviewed below highlight both the advances made and the persistent gaps that motivate the present work.

A. Encryption in Resource-Constrained IoT Environments

Lightweight cryptographic approaches have been proposed to address the resource constraints of IoT devices. Rahman et al. [14] proposed chaos-based AES key generation to strengthen symmetric encryption for smart home devices, demonstrating improved resistance to brute-force attacks while retaining computational feasibility. Tripathy and Singh [8] evaluated AES software implementations on IoT-class hardware, finding that standard AES configurations require optimization to meet real-time latency requirements on constrained processors. Fernandez et al. [9] conducted a comparative evaluation of RSA and ECC cipher suites on fog and mist computing devices, concluding that ECC provides a more favorable security-to-overhead ratio for IoT applications, though it introduces key management and side-channel considerations [10].

Patil et al. [2] proposed a hybrid adaptive cryptographic authentication approach that selects encryption algorithms based on device context, demonstrating that adaptive selection can reduce computational overhead without compromising security guarantees. These works collectively establish that no single cryptographic primitive satisfies the full range of IoT security requirements, motivating layered and adaptive approaches.

B. Anomaly Detection and Intrusion Detection in IoT

Khraisat and Alazab [12] provide a comprehensive review of intrusion detection techniques for IoT environments, highlighting that signature-based methods fail against novel attack patterns and that machine learning approaches, particularly unsupervised methods, offer broader detection coverage. Hamarsheh [15] proposed an adaptive security framework combining Software-Defined Networking (SDN) and machine learning for dynamic IoT threat response, demonstrating measurable improvement in detection latency and policy enforcement speed relative to static rule-based systems. Anthi et al. [17] developed EclipseIoT, a hub-based adaptive security system that monitors device behavior and enforces isolation policies in response to detected anomalies, providing a concrete point of comparison for hub-centric adaptive frameworks.

Both Hamarsheh [15] and EclipseIoT [17] address adaptive detection but do not integrate layered cryptographic policy enforcement across device, network, and cloud tiers simultaneously. This distinction is central to the contribution of AMLE.

C. Access Control in IoT Cloud Environments

Attribute-Based Access Control (ABAC) has been studied as a flexible mechanism for IoT access management. Pajooh et al. [16] proposed a multi-layer blockchain-based security architecture that incorporates ABAC-style policies for IoT data access, demonstrating that attribute-driven authorization can be enforced at the cloud layer without requiring device-side computation. Mosenia and Jha [5] identify weak access control as a primary vulnerability in deployed IoT systems, noting that most consumer devices lack configurable authorization mechanisms, underscoring the importance of cloud-tier access control solutions that operate independently of device capabilities.

D. Identified Research Gap

The reviewed literature demonstrates that encryption optimization, anomaly detection, and access control have each been studied in isolation or in partial combinations. However, a unified framework that coordinates adaptive encryption policy enforcement across device, network, and cloud layers in response to real-time threat signals remains absent from the literature. Existing adaptive frameworks such as Hamarsheh [15] and EclipseIoT [17] address detection and isolation but do not systematically couple detection outcomes to multi-level cryptographic responses. The AMLE model presented in this study addresses this gap by integrating lightweight device-layer encryption, machine learning-based network anomaly detection, and cloud-layer ABAC within a unified policy-driven architecture.

III. Methodology

A. Research Design

The study adopted a prototype development methodology using an Agile approach with iterative development cycles. This approach supported incremental design, implementation, and validation of AMLE components through successive sprints. The development process comprised four phases: planning, sprint execution, review, and deployment.

During the planning phase, key security challenges in smart home IoT ecosystems were identified, including limited device resources, diverse communication protocols, and scalability requirements [3]. These constraints informed design decisions, particularly the selection of encryption mechanisms and anomaly detection techniques compatible with resource-constrained environments.

B. System Architecture

The AMLE architecture follows a three-layer IoT security model aligned with established IoT cloud environment frameworks [13]. Each layer addresses distinct security requirements while contributing to an integrated security posture. Fig. 1 illustrates the overall system architecture showing security enforcement across device, network, and cloud layers.

1) Device layer: At the device layer, lightweight encryption mechanisms including ECC and AES-128 are employed to protect data at the source while accommodating device resource constraints [8]. Encryption strength is assigned based on predefined policies that consider device computational capacity and data sensitivity [14]. For example, devices handling sensitive data such as video streams apply stronger encryption compared to low-risk devices such as temperature sensors. This policy-driven differentiation reflects the finding that cryptographic mechanisms must be tailored to device capabilities and application context in IoT environments [2].

2) Network layer: The network layer incorporates anomaly detection using the Isolation Forest machine learning algorithm to analyze traffic patterns and identify deviations from established baseline behavior [12]. When anomalous activity is detected, predefined response policies may trigger actions such as increased encryption strength, connection restriction, or alert generation. This approach addresses limitations of traditional

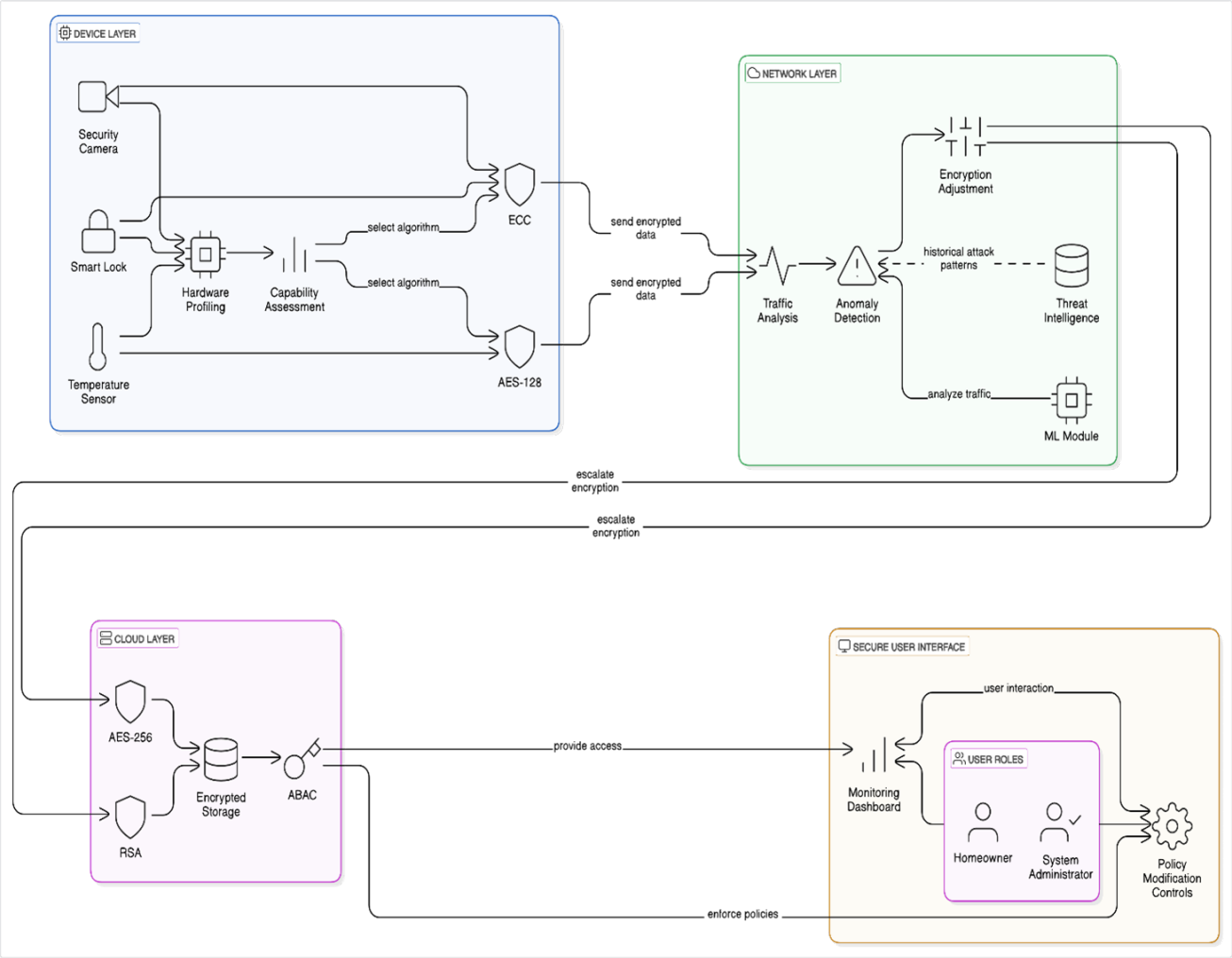


Fig. 1. Overall AMLE system architecture illustrating security enforcement across device, network, and cloud layers.

signature-based intrusion detection systems by enabling detection of previously unseen patterns without requiring labeled training data.

3) Cloud layer: The cloud layer applies robust encryption mechanisms including AES-256 and RSA for data at rest and in transit, combined with Attribute-Based Access Control (ABAC) to enforce fine-grained access policies. ABAC decisions are based on user attributes, resource characteristics, and environmental conditions, providing flexible and context-aware access control.

This approach is consistent with findings from Pajooh et al. [16], who demonstrate that attribute-driven authorization is computationally feasible at the cloud tier of IoT architectures and can be enforced without imposing additional processing burdens on resource-constrained edge devices.

C. Implementation Environment

The system was implemented using a software stack selected to support cryptographic operations, machine learning, and scalable API services. Table I summarizes the implementation environment and toolchain used throughout the development and testing phases.

TABLE I. IMPLEMENTATION ENVIRONMENT AND TOOLCHAIN

Component	Technology / Specification
Programming Language	Python
API Framework	FastAPI v0.104.1
Machine Learning Library	Scikit-learn v1.3.0
Cryptographic Library	Cryptography v41.0.5
Database ORM	SQLAlchemy v2.0.23
Hardware Platform	Intel i7 CPU, 16 GB RAM, SSD
Deployment Context	Controlled simulation environment

This environment supported development and validation of encryption operations, anomaly detection, and access control logic under repeatable test conditions. Testing was conducted on an Intel i7 processor with 16 GB RAM and SSD storage.

IoT devices were simulated in software to represent typical smart home devices with varying processing capabilities and security requirements [17]. No physical IoT hardware was used; all device behavior, traffic generation, and threat

TABLE II. ISOLATION FOREST CONFIGURATION PARAMETERS

Parameter	Value	Rationale
Algorithm	Isolation Forest	Unsupervised anomaly detection
Number of Estimators	100	Balance between accuracy and efficiency
Contamination Rate	0.05	Consistent with IoT anomaly proportion estimates [12]
Feature Normalization	StandardScaler	Prevent feature dominance
Detection Frequency	Every 15 min	Periodic behavioral analysis

scenarios were produced within the simulation environment. Performance and security results therefore reflect behavior under controlled simulation conditions, and generalization to bare-metal IoT hardware requires further empirical validation on representative physical platforms.

D. Anomaly Detection Implementation

Isolation Forest was selected due to its suitability for unsupervised anomaly detection in high-dimensional datasets without requiring labeled training data [12]. Feature selection included traffic volume metrics, connection characteristics, and encryption-related attributes such as algorithm type and key rotation frequency. Table II presents the configuration parameters used for the Isolation Forest implementation.

The contamination rate of 0.05 was selected based on reported estimates of anomalous traffic proportions in IoT network studies [12]. It is acknowledged that this value was not empirically derived from a dedicated IoT attack dataset in this study; sensitivity to this parameter is noted as a limitation, and future work should calibrate the contamination rate against established IoT traffic benchmarks such as N-BaIoT and IoT-23 to better characterize detection performance across varying threat densities.

Data preprocessing applied normalization using StandardScaler, and traffic data were aggregated into fixed time intervals to capture behavioral patterns over time. Anomaly detection was executed at scheduled intervals, with detected anomalies logged and assigned severity levels for monitoring and response.

E. Encryption Management

AMLE defines four encryption levels (LOW, MEDIUM, HIGH, and VERY HIGH) based on algorithm selection, key size, and rotation frequency [15]. Table III details the characteristics of each encryption level and their intended device profiles.

Devices are assigned encryption levels according to predefined policies considering device capabilities, data sensitivity, and detected threat indicators. Key management includes per-device key generation and rotation, implemented using time-based schedules and threat-triggered events. Supported algorithms include AES, RSA, and ECC, enabling flexibility across heterogeneous IoT devices [9], [10].

F. Testing Methodology

Testing comprised unit, integration, functional, performance, and security testing phases. Unit tests verified encryption operations, key management functions, and policy enforcement mechanisms. Integration testing validated secure data flow across layers. Performance testing evaluated encryption latency, system scalability, and resource utilization under varying simulated loads. Security testing included penetration testing and threat modeling to assess system behavior under representative attack scenarios.

It is noted that all threat scenarios were generated internally within the simulation environment; evaluation against established IoT attack datasets such as those reviewed by Khraisat and Alazab [12] represents an important direction for future validation, as externally sourced datasets would provide a more rigorous basis for assessing detection coverage and false positive rates.

IV. RESULTS

A. Key Management System

Key management tests confirmed correct execution of key lifecycle operations, including automated rotation, state-aware updates, and audit trail generation. Invalid configurations were rejected, and inactive devices were excluded from key updates, demonstrating correct policy enforcement and system robustness. Table IV summarizes the functional test outcomes for the key management system.

B. Encryption and Decryption

Encryption and decryption processes functioned as intended across all tested scenarios. Encrypted data could not be accessed without valid keys, and decrypted data matched original input, confirming data confidentiality and integrity. Multi-layer encryption remained consistent across device, network, and cloud layers. Table V presents validation results for encryption and decryption operations.

C. System Performance

Encryption and decryption operations completed within latency ranges suitable for typical smart home interactions, such as sensor updates and control commands. The system handled concurrent simulated device connections without observable degradation in response time. Resource utilization remained within acceptable limits under the tested conditions, although increased latency was observed for configurations representing highly constrained devices.

Table VI summarizes the observed performance characteristics. Performance measurements are reported qualitatively because the evaluation was conducted in a software simulation environment on a general-purpose workstation rather than on physical IoT hardware; precise numerical latency benchmarks on constrained devices remain a subject for future empirical study, and direct measurement on representative hardware platforms will be necessary to establish deployment-ready performance profiles.

D. Anomaly Detection

The Isolation Forest model identified deviations from baseline traffic behavior, including abnormal transmission volumes, unusual connection patterns, and inconsistent encryption parameters. Detected anomalies were logged with severity

TABLE III. ENCRYPTION LEVELS AND CHARACTERISTICS

Level	Algorithms	Key Size	Rotation Policy	Intended Device Profile
LOW	AES	128-bit	Periodic	Low-risk sensors
MEDIUM	AES	192-bit	Periodic	Standard IoT devices
HIGH	AES / ECC	256-bit	Periodic + Threat-triggered	Sensitive devices
VERY HIGH	AES / RSA	256-bit	Frequent + Threat-triggered	Cloud and critical assets

TABLE IV. KEY MANAGEMENT FUNCTIONAL TEST OUTCOMES

Test Case	Expected Behavior	Outcome
Automatic key rotation	Rotation at scheduled intervals	Successful
Invalid interval handling	Reject invalid values	Successful
Inactive device handling	Exclude from updates	Successful
Rotation audit logging	Retain rotation history	Successful
Encryption disabled state	Suspend rotation	Successful

TABLE V. ENCRYPTION AND DECRYPTION VALIDATION RESULTS

Scenario	Expected Outcome	Result
Correct key usage	Successful decryption	Confirmed
Incorrect key usage	Decryption failure	Confirmed
Multi-layer encryption	End-to-end protection	Confirmed
Data integrity check	Original data restored	Confirmed

TABLE VI. OBSERVED PERFORMANCE CHARACTERISTICS

Metric	Observation
Encryption latency	Within acceptable range for simulated smart home interactions
Concurrent device handling	Stable under simulated load
CPU utilization	Moderate under peak conditions
Memory usage	Within available limits

TABLE VII. DETECTED ANOMALY CATEGORIES

Category	Description
Traffic volume anomalies	Unexpected data transmission spikes
Connection pattern anomalies	Abnormal frequency or duration
Encryption anomalies	Deprecated algorithms or irregular key rotation
Destination anomalies	Connections to unknown or blacklisted IPs

ratings and visualized via the monitoring dashboard, enabling timely security responses. Table VII categorizes the types of anomalies detected during testing.

E. Security Behavior Under Simulated Threats

The framework demonstrated the ability to detect and respond to simulated unauthorized access attempts, MitM-like traffic interception patterns, and botnet-associated communication behaviors. Cloud-layer ABAC policies restricted unauthorized access attempts based on attribute mismatches, supporting effective access control enforcement. Table VIII summarizes the security response outcomes under various threat scenarios.

All threat scenarios were generated internally within the simulation environment; validation against externally sourced IoT attack datasets is acknowledged as a limitation and is identified as a direction for future work, as such validation would provide stronger evidence of the framework's generalizability to real-world attack distributions and previously unseen threat variants.

TABLE VIII. SECURITY RESPONSE SUMMARY

Threat Scenario	Detection	Response Mechanism
Unauthorized access attempt	Detected	Access blocked via ABAC
MitM-like traffic pattern	Detected	Encryption escalation
Botnet-style behavior	Detected	Alert and connection restriction
Unauthorized cloud access	Detected	Policy-based denial

V. DISCUSSION

The AMLE framework demonstrates that layered, policy-driven security mechanisms can enhance protection in smart home IoT environments while accommodating resource constraints. Lightweight encryption at the device layer provides baseline protection, while network-level anomaly detection enables adaptive responses to detected deviations.

A. Device Layer Performance

The device layer implementation of ECC and AES-128 provided data security at the source while accommodating computational constraints characteristic of IoT devices [8], [10]. Dynamic adjustment of encryption strength based on device capabilities and data sensitivity optimized the balance between security and performance [14]. The differential encryption approach, where security cameras utilize stronger protocols than temperature sensors, is consistent with the adaptive selection approach demonstrated by Patil et al. [2], who show that context-driven algorithm selection reduces overhead without degrading security guarantees.

B. Network Layer Effectiveness

The network layer machine learning-based anomaly detection successfully identified suspicious traffic patterns, enabling real-time encryption escalation when threats were detected. The Isolation Forest unsupervised learning approach proved effective for detecting previously unseen attack patterns without requiring labeled training data, addressing a critical gap in conventional intrusion detection systems [12]. The configuration with 100 estimators and a 0.05 contamination threshold balanced detection sensitivity with computational efficiency.

Direct quantitative comparison with Hamarsheh [15] and EclipseIoT [17] on shared detection metrics was not performed in this study because the evaluation environments and datasets differ; such comparison on common benchmarks is identified as a priority for future work and would substantially strengthen

claims regarding the relative performance of AMLE's anomaly detection component.

C. Cloud Layer Security

The cloud layer implementation of AES-256, RSA encryption, and ABAC mechanisms ensured secure data storage and restricted unauthorized access [16]. The combination of robust encryption algorithms and attribute-based access control provided comprehensive protection for sensitive information.

ABAC's ability to make decisions based on user attributes, resource characteristics, and environmental conditions enabled flexible and context-aware security enforcement, consistent with findings from Pajooh et al. [16] in multi-layer IoT architectures. This flexibility is particularly valuable in smart home contexts where access permissions may need to vary dynamically based on factors such as time of day, device health status, or detected threat level.

D. Performance and Scalability

Performance results indicate that the framework is suitable for simulated smart home scenarios, though the simulation on a general-purpose Intel i7 workstation does not directly confirm feasibility on physically constrained IoT hardware. Encryption and decryption operations completed within latency ranges suitable for interactive smart home functions such as door lock commands, thermostat adjustments, and security alerts. The system scaled effectively to accommodate multiple simultaneous device connections under simulation.

Increased latency was observed for configurations representing highly constrained devices, consistent with findings from studies on cryptographic implementation in minimal IoT devices [8].

This limitation indicates that further optimization of encryption algorithms for physically constrained devices, potentially through hardware acceleration or more efficient algorithmic implementations, is necessary before deployment on actual IoT hardware can be claimed. Future iterations of AMLE should therefore incorporate profiling on representative low-power platforms such as Raspberry Pi Zero or Arduino-class microcontrollers to quantify the overhead imposed by each encryption level.

E. System Integration

The modular architecture supports extensibility and integration with existing IoT infrastructures. The scheduled anomaly detection executing every 15 minutes, combined with database integration and RESTful API endpoints, provides practical security monitoring that can be integrated into existing smart home management systems. The dashboard implementation offers administrators comprehensive visibility into encryption status, device health, and security metrics, enabling proactive security management and rapid response to detected threats.

VI. Conclusion

This study presented an Adaptive Multi-Layer Encryption framework for smart home IoT ecosystems that integrates lightweight cryptography, network-level anomaly detection, and cloud-based access control. Evaluation in a controlled simulation environment demonstrated the framework's capability to detect and respond to representative security threats while maintaining operational performance suitable for typical smart home applications.

It is explicitly acknowledged that the evaluation was conducted entirely in software simulation on a general-purpose workstation and that performance and detection claims require further validation on physical IoT hardware and against established IoT attack datasets. These limitations define clear boundaries on the current findings and motivate the empirical extensions outlined in the future work directions below.

The research contributes to the field of IoT security by providing a practical encryption framework that balances security requirements with operational constraints of smart home environments. The validated implementation of multi-layer adaptive encryption across device, network, and cloud layers establishes a reference architecture for future smart home security systems. Integration of machine learning-based anomaly detection for real-time threat response demonstrates the feasibility of intelligent, adaptive security mechanisms in simulated resource-constrained environments.

A. Key Contributions

The findings support the feasibility of layered, adaptive security architectures for IoT environments and provide a reference implementation for future research and development. Key contributions include a practical encryption framework validated through comprehensive testing, demonstration of anomaly detection using Isolation Forest with a 0.05 contamination threshold calibrated to IoT traffic estimates in the literature [12], and evidence of acceptable performance characteristics for typical smart home applications under controlled simulation conditions.

B. Future Work

Future work should focus on: validation in real-world deployments on physical IoT hardware; calibration of the anomaly detection contamination parameter against established IoT attack datasets such as N-BaIoT and IoT-23; quantitative benchmarking against adaptive IoT security frameworks such as Hamarsheh [15] and EclipseIoT [17] on shared metrics; optimization for highly constrained devices via hardware acceleration; expanded protocol support for broader compatibility; integration with blockchain for distributed trust, enabling immutable audit trails and non-repudiation through digital signature schemes [18]; and development of standardized AMLE protocols for industry adoption.

The AMLE model provides a foundation for developing more robust, adaptive security solutions for the expanding smart home IoT ecosystem, contributing to safer and more reliable deployment of IoT technologies in residential environments.

Declaration on Generative AI Use

The authors declare that generative artificial intelligence (GenAI) tools were used solely for language refinement, grammar checking, and improving clarity and readability of the

manuscript. The scientific content, methodology, experimental design, analysis, results, and conclusions were entirely developed by the human authors. All AI-assisted text was carefully reviewed, verified, and edited to ensure accuracy, originality, and compliance with academic integrity standards. The authors take full responsibility for the content of this manuscript.

REFERENCES

[1] A. Chakraborty, M. Islam, F. Shahriyar, S. Islam, H. U. Zaman, and M. Hasan, "Smart home system: A comprehensive review," *J. Electr. Comput. Eng.*, vol. 2023, Article ID 7616683, 2023.

[2] K. Patil, I. Mandal, and C. Rangaswamy, "Hybrid and adaptive cryptographic-based secure authentication approach in IoT-based applications using hybrid encryption," *Pervasive Mobile Comput.*, vol. 82, p. 101552, 2022.

[3] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, "Security, privacy and trust in Internet of Things: The road ahead," *Comput. Networks*, vol. 76, pp. 146–164, Jan. 2015.

[4] W. Z. Khan, M. Y. Aalsalem, M. K. Khan, and Q. Arshad, "Data and privacy: Getting consumers to trust products enabled by the Internet of Things," *IEEE Consum. Electron. Mag.*, vol. 8, no. 2, pp. 35–38, Mar. 2019.

[5] A. Mosenia and N. K. Jha, "A comprehensive study of security of Internet-of-Things," *IEEE Trans. Emerging Topics Comput.*, vol. 5, no. 4, pp. 586–602, Oct.–Dec. 2017.

[6] P. Radanliev, D. De Roure, K. Page, J. R. C. Nurse, R. M. Montalvo, O. Santos, L. Maddox, and P. Burnap, "Cyber risk at the edge: Current and future trends on cyber risk analytics and artificial intelligence in the industrial Internet of Things and Industry 4.0 supply chains," *Cybersecurity*, vol. 3, no. 1, p. 13, Jun. 2020.

[7] R. Das and M. Z. Gündüz, "Analysis of cyber-attacks in IoT-based critical infrastructures," *Int. J. Inf. Secur. Sci.*, vol. 8, no. 4, pp. 122–133, 2019.

[8] A. Tripathy and B. Singh, "A study of AES software implementation for IoT systems," in *Proc. 3rd Int. Conf. Issues Challenges Intell. Comput. Techn. (ICICT)*, Feb. 2022, pp. 1–4.

[9] T. Fernández-Caramés, P. Fraga-Lamas, and M. Suárez-Albela, "A practical evaluation on RSA and ECC-based cipher suites for IoT high-security energy-efficient fog and mist computing devices," *Sensors*, vol. 18, no. 11, p. 3868, Nov. 2018.

[10] S. Ullah, J. Zheng, N. Din, M. T. Hussain, F. Ullah, and M. Yousaf, "Elliptic curve cryptography: Applications, challenges, recent advances, and future trends: A comprehensive survey," *Comput. Sci. Rev.*, vol. 47, p. 100530, Feb. 2023.

[11] O. Alrawi, C. Lever, M. Antonakakis, and F. Monrose, "SoK: Security evaluation of home-based IoT deployments," in *Proc. IEEE Symp. Security Privacy (SP)*, May 2019, pp. 1362–1380.

[12] A. Khraisat and A. Alazab, "A critical review of intrusion detection systems in the Internet of Things: Techniques, deployment strategy, validation strategy, attacks, public datasets and challenges," *Cybersecurity*, vol. 4, no. 1, p. 18, May 2021.

[13] A. Bouaouad, A. Cherradi, S. Assoul, and N. Souissi, "The key layers of IoT architecture," in *Proc. 5th Int. Conf. Cloud Comput. Artif. Intell.: Technol. Appl. (CloudTech)*, Nov. 2020, pp. 1–4.

[14] Z. Rahman, X. Yi, M. Billah, M. Sumi, and A. Anwar, "Enhancing AES using chaos and logistic map-based key generation technique for securing IoT-based smart home," *Electronics*, vol. 11, no. 7, p. 1083, Apr. 2022.

[15] A. Hamarsheh, "An adaptive security framework for Internet of Things networks leveraging SDN and machine learning," *Appl. Sci.*, vol. 14, no. 11, p. 4530, May 2024.

[16] H. Honar Pajooh, M. Rashid, F. Alam, and S. Demidenko, "Multi-layer blockchain-based security architecture for Internet of Things," *Sensors*, vol. 21, no. 3, p. 772, Jan. 2021.

[17] E. Anthi, S. Ahmad, O. Rana, G. Theodorakopoulos, and P. Burnap, "EclipseIoT: A secure and adaptive hub for the Internet of Things," *Comput. Security*, vol. 78, pp. 477–490, Sep. 2018.

[18] W. Fang, W. Chen, W. Zhang, J. Pei, W. Gao, and G. Wang, "Digital signature scheme for information non-repudiation in blockchain: A state-of-the-art review," *EURASIP J. Wireless Commun. Networking*, vol. 2020, p. 56, 2020.

Workload-Aware Storage Reduction for Multi-Tenant SIEM on ClickHouse

Nutthakorn Chalaemwongwan
Department of Computer Engineering, KOSEN-KMITL,
King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

***Abstract*—Security Information and Event Management (SIEM) platforms ingest terabytes of heterogeneous telemetry daily—Windows event logs, DNS queries, HTTP transactions, EDR alerts, and network metadata from Zeek—yet the majority of stored records are never queried for threat-hunting or incident-response workflows. This study presents a workload-aware storage reduction framework that tailors data retention to observed analytical demand within a multi-tenant ClickHouse deployment. The main contribution is a Workload Analyzer algorithm that extracts importance scores for columns from ClickHouse query logs using a frequency–recency–coverage weighting scheme, and a Storage-Coverage Cost Model that computes the optimal pruning threshold that minimizes a weighted sum of storage cost and coverage loss. Guided by these metrics, the framework applies six composable reduction operators: column pruning with materialized views, adaptive sampling, deduplication, per-column codec selection, skip-indexing, and time-to-live (TTL)-based retention tiering across hot/warm/cold storage. Multi-tenant isolation is enforced through role-based access control overlays aligned with the Thai Personal Data Protection Act (PDPA). Experimental evaluation on 1,000,000 Zipf-distributed Windows Security Events demonstrates 79% uncompressed and 70% compressed storage reduction with sub-second query latency, while the Workload Analyzer automatically identifies the optimal column subset that preserves 100% detection rule coverage at minimum storage cost.**

Keywords*—ClickHouse; SIEM; storage reduction; workload analysis; column importance scoring; multi-tenant; PDPA compliance

I. Introduction

Every SOC has a SIEM and every SIEM has storage. A mid-sized enterprise generates 200–500 GB of security telemetry per day, and big shops do 1 TB or more [1]–[3]. At cloud prices, this translates to tens of thousands of dollars a month for storage with no end in sight [4].

Splunk billing is by gigabytes ingested, so analysts are on the lookout for what they send to the platform [5], [6]. Switching to an open-source stack like ClickHouse [7] eliminates the costly licenses but does not address the fact that the volume of data is not slowing down.

The ugly reality we kept running into was that most of the bytes never get queried. With an examination of real-world query logs, we found that only a small set of columns matter for Sigma rules [8], [9] and for investigation playbooks. The rest of the data is dead weight burning the budget without raising a single alert [3].

We propose a **workload-aware storage reduction framework** for multi-tenant SIEM platforms on ClickHouse. Unlike prior work that applies individual techniques in isolation—compression here, sampling there—our approach *composes* six reduction operators into a single pipeline whose pruning decisions are driven by a formal cost model that explicitly balances storage savings against detection coverage loss. This composition, guided by workload-aware optimization, is what distinguishes our work from both database-side automatic physical design [10], [11] and SIEM-side lifecycle management [1], [2].

We make four contributions:

- A Workload Analyzer that mines ClickHouse query logs to score every column by frequency, recency, and detection rule coverage—then automatically decides what to keep and what to drop.
- A Storage-Coverage Cost Model that frames column pruning as an explicit optimization: minimize storage cost subject to a penalty for losing detection coverage.
- A composable pipeline of six reduction operators—column pruning, adaptive sampling, deduplication, codec selection, skip-index acceleration, and TTL-tiered retention—engineered for security telemetry on ClickHouse.
- A multi-tenant infrastructure that supports the Thai PDPA [12], achieving 79% raw data storage reduction with sub-second query latency.

The study proceeds as follows: Section II surveys related work. Section III describes the system architecture. Section IV presents the Workload Analyzer. Section V and Section VI cover the reduction operators and tenant isolation, respectively. Section VII reports experiments. Section VIII and Section IX discuss findings and conclusions.

II. Related Work

A. SIEM Platforms and Log Management

Mainstream SIEM products (e.g., as reviewed by Bhatt et al. [13])—Splunk [1], IBM QRadar [14], Elastic Security [2]—all do log management, but at almost-doubled storage cost. Splunk extracts fields at index time and keeps raw payload around; QRadar discards part of raw data in its normalized event model. Net effect is to store roughly double size of original log files on disk.

Open-source deployments like Wazuh [15] and Security Onion [16] fall back on Elasticsearch, which stores data in

an inverted-index format. The ILM features of Elasticsearch enable rolling indices and scheduled deletion, but this is a time-based, waveform-pattern approach, not a workload-aware column-level optimization strategy. The distinction is important.

B. Columnar Storage for Security Analytics

Columnar databases are being used for security analytics. ClickHouse [7], [17], which was initially developed at Yandex for web analytics, provides native columnar storage with high-compression ratios, making it a popular choice for log analysis. The ClickHouse VLDB industrial paper [17] shows that columnar storage with per-column codecs significantly outperforms row-oriented storage for log analysis, and the Huntress team showed that it is feasible to build a complete SIEM on ClickHouse even at a scale of billions of events per day [4] (complemented by recent academic evaluations of cloud-native log storage [6], [18]). Columnar storage for log analysis for cloud-native security information and event management (SIEM) is also affordable in the cloud, as shown by Zhu et al. [6].

Column-oriented storage in general allows for column pruning (only reading columns that are used in a query from disk), but only at query time, and the entire schema is still stored. Our approach extends this to prune unused columns at ingestion time using materialized views, which removes storage overhead permanently.

C. Data Reduction Techniques

Data reduction for log management has been studied in several contexts. Sampling approaches [19] reduce data volume but risk losing rare, security-critical events. Deduplication [20] is effective for repetitive logs but requires careful handling of event ordering and provenance. Compression-based approaches exploiting domain-specific patterns [21] have shown significant gains in analytical databases. Recent work on log compression by Li et al. [18] demonstrates that exploiting both static and runtime patterns can substantially reduce cloud log storage costs. He et al. [3] provide an empirical study of log parsing practices in industry.

Query-driven workload analysis for database optimization has been studied by Ma et al. [10], who proposed workload forecasting for self-driving databases, and more recently by Perera et al. [22], who developed adaptive learned query optimizers for multi-tenant systems. Automatic materialized view generation using deep learning has been explored by Liu et al. [11]. However, none of these approaches target the specific combination of security telemetry, detection rule coverage preservation, and SIEM workload patterns that our framework addresses.

Our work differs from prior approaches by *composing* multiple reduction techniques into a unified, workload-aware pipeline that is specifically designed for security telemetry and multi-tenant deployment. Table I summarizes the key differences.

D. Multi-Tenancy and Data Privacy

Multi-tenant database architectures have been extensively studied [23], [24]. In the context of security platforms, tenant isolation carries additional requirements due to the sensitivity of security telemetry [22]. The Thai PDPA [12], modeled after the EU GDPR [25], [26], mandates purpose limitation, data minimization, and storage limitation—principles that align naturally with workload-aware storage reduction.

TABLE I. COMPARISON WITH EXISTING APPROACHES

Approach	Workload Aware	Column Pruning	Detection Coverage	Multi-Tenant	Privacy (PDPA)
Splunk ILM [1]	–	–	–	✓	–
Elastic ILM [2]	–	–	–	✓	–
ClickHouse [17]	–	✓	–	–	–
Huntress [4]	–	–	–	✓	–
Ma et al. [10]	✓	–	–	–	–
Liu et al. [11]	✓	✓	–	–	–
Ours	✓	✓	✓	✓	✓

III. SYSTEM ARCHITECTURE

Fig. 1 shows the overall system. Four subsystems work together: an ingestion layer, a reduction operator pipeline, a multi-tenant overlay, and a tiered storage backend. We walk through each below.

A. Ingestion Layer

The ingestion layer accepts security telemetry from heterogeneous sources, including endpoint agents (Beats, Wazuh), EDR platforms, Windows Event Forwarding (WEF), DNS resolvers, HTTP proxies, and Zeek network sensors. All events are serialized as JSON and published to Apache Kafka [27] topics partitioned by source type and tenant identifier.

Kafka works as a strong abstraction between event producers and downstream storages. That matters in practice: during a security incident, a spike in burst ingestion can bring ingestion rates $10\times$ above baseline, and Kafka absorbs that without backing pressure against the agents.

B. Reduction Operator Pipeline

Events consumed from Kafka undergo a parameterizable 6-stage reduction pipeline (Section V), applying the operations in an order that maximizes cumulative reduction:

- Column pruning (dropping unused columns)
- Deduplication and canonicalization (merging duplicates)
- Adaptive sampling (downsampling low-value event types)
- Codec selection (best compression strategy per column)
- Skip-index build (supporting frequent match-all scans)
- TTL assignment (applying tiered expiration policies)

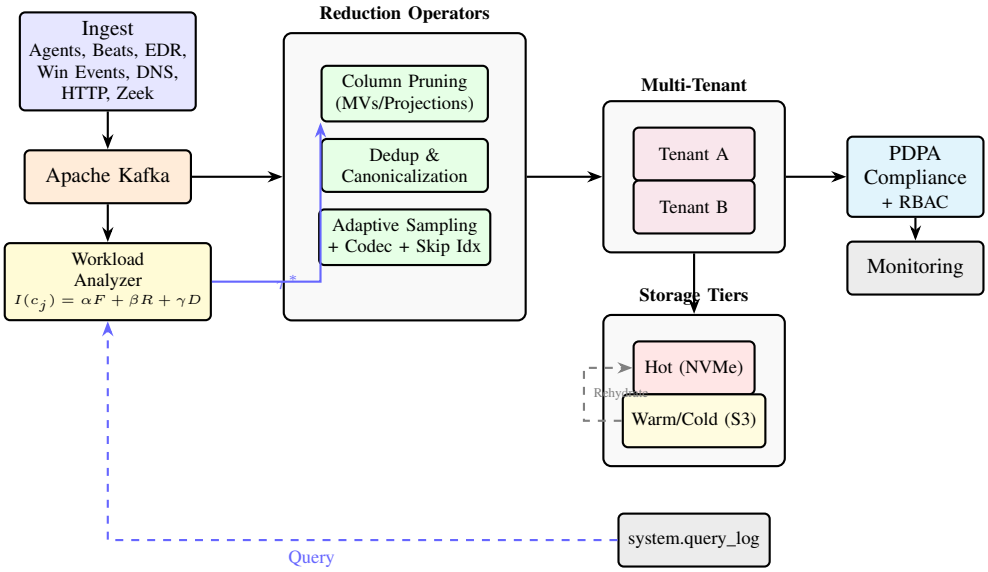


Fig. 1. System architecture of the workload-aware storage reduction framework. Security telemetry flows from diverse sources through Kafka into the reduction operator pipeline, which applies six composable operators before storing data in a multi-tenant, tiered ClickHouse deployment. RBAC overlays enforce tenant isolation and PDPA compliance.

C. Multi-Tenant Overlay

The multi-tenant overlay provides logical isolation between tenants within a shared ClickHouse cluster. Each tenant's data is tagged with a `tenant_id` column and access is enforced through ClickHouse's row-level security policies and role-based access control (RBAC). Section VI details this subsystem.

D. Tiered Storage Backend

ClickHouse's native storage tiering is configured with three tiers: **Hot** (NVMe SSD) for the most recent and frequently queried data, **Warm** (HDD or network-attached) for aging data, and **Cold** (S3-compatible object storage) for long-term archival. TTL rules automatically migrate data between tiers based on age, and on-demand rehydration allows cold data to be temporarily promoted to hot storage for investigation workflows.

IV. Workload Analyzer and Cost Model

How do we decide which columns to keep? Doing this by hand is tedious and breaks whenever someone adds a new Sigma rule [9]. Previous work on query workload forecasting [10] and learned indexes [28] has shown that mining query logs is a viable path to automation. We take that idea further: this section describes an algorithm that scores every column and a cost model that picks the best pruning threshold.

A. Column Importance Scoring

The Workload Analyzer continuously mines ClickHouse's `system.query_log` to extract column access patterns. For each column c_j in the schema, it computes an importance score $I(c_j)$ [using Eq. (1)] that reflects three dimensions:

$$I(c_j) = \alpha \cdot F(c_j) + \beta \cdot R(c_j) + \gamma \cdot D(c_j) \tag{1}$$

where, $\alpha + \beta + \gamma = 1$ and:

- $F(c_j) = \frac{n(c_j)}{N}$ is the **frequency score**: the fraction of queries in the observation window that reference column c_j, where $n(c_j)$ is the number of queries referencing c_j and N is the total query count.
- $R(c_j) = \frac{1}{M} \sum_{i=1}^{M} e^{-\lambda \cdot \Delta t_i}$ is the **recency score**: an exponentially-weighted average of the time elapsed Δt_i (in days) since each of the M most recent accesses to c_j, with decay parameter $\lambda > 0$.
- $D(c_j) = \frac{|\mathrm{R}(c_j)|}{|\mathrm{R}|}$ is the **detection coverage score**: the fraction of active Sigma detection rules R that reference column c_j.

The frequency score captures routine analytical demand, the recency score ensures that recently active columns are preserved even if their long-term frequency is low (e.g., during an active investigation), and the detection coverage score prevents pruning of columns that are critical to threat detection rules.

B. Automatic Column Classification

Algorithm 1 presents the complete Workload Analyzer procedure. Given a pruning threshold τ and the schema S, it classifies each column as RETAIN (include in the pruned materialized view) or PRUNE (exclude from long-term storage).

The protected set P contains columns that must never be pruned regardless of their importance score (e.g., primary keys, tenant identifiers, timestamps). This ensures structural integrity of the ClickHouse MergeTree engine.

C. Storage-Coverage Cost Model

The pruning threshold τ controls the trade-off between storage savings and detection coverage. We formulate this as

Algorithm 1 Workload Analyzer: Column Classification

Require: Schema $\mathrm{S} = \{c_1, c_2, \ldots, c_n\}$, threshold τ, weights (α, β, γ), decay λ, window W days, protected set P
Ensure: Classification $\mathrm{C} : c_j \mapsto \{\textsc{Retain}, \textsc{Prune}\}$
$\mathrm{Q} \leftarrow \textsc{QueryLog}(W)$ ▷ Fetch queries from last W days
$\mathrm{R} \leftarrow \textsc{LoadSigmaRules}$ ▷ Active detection rules
$N \leftarrow |\mathrm{Q}|$
for each column $c_j \in \mathrm{S}$ **do**
  $n_j \leftarrow |\{q \in \mathrm{Q} : c_j \in \mathrm{cols}(q)\}|$
  $F_j \leftarrow n_j \,/\, N$
  $\{\Delta t_1, \ldots, \Delta t_M\} \leftarrow \textsc{RecentAccessTimes}(c_j, M)$
  $R_j \leftarrow \frac{1}{M} \sum_{i=1}^{M} e^{-\lambda \cdot \Delta t_i}$
  $D_j \leftarrow \left|\{r \in \mathrm{R} : c_j \in \mathrm{cols}(r)\}\right| / |\mathrm{R}|$
  $I_j \leftarrow \alpha \cdot F_j + \beta \cdot R_j + \gamma \cdot D_j$
  if $c_j \in \mathrm{P}$ **or** $I_j \geq \tau$ **then**
    $\mathrm{C}(c_j) \leftarrow \textsc{Retain}$
  else
    $\mathrm{C}(c_j) \leftarrow \textsc{Prune}$
  end if
end for
return C

an optimization problem.

Let $s(c_j)$ denote the average storage cost (bytes per row) of column c_j, and let $\mathrm{S}_\tau = \{c_j : I(c_j) \geq \tau\} \cup \mathrm{P}$ be the set of retained columns at threshold τ. The total storage cost is given by Eq. (2):

$$C_{\text{storage}}(\tau) = \sum_{c_j \in \mathrm{S}_\tau} s(c_j) \tag{2}$$

The detection coverage at threshold τ measures what fraction of active detection rules remain fully executable structurally, as defined in Eq. (3). Note that this metric counts only whether the required columns are retained, not whether individual events themselves survive adaptive sampling:

$$\mathrm{Cov}(\tau) = \frac{|\{r \in \mathrm{R} : \mathrm{cols}(r) \subseteq \mathrm{S}_\tau\}|}{|\mathrm{R}|} \tag{3}$$

The optimal threshold τ^* minimizes the combined cost formulated in Eq. (4):

$$\tau^* = \arg\min_{\tau \in [0,1]} \underbrace{\frac{C_{\text{storage}}(\tau)}{C_{\text{storage}}(0)}}_{\text{normalized storage}} + w \cdot \underbrace{\big(1 - \mathrm{Cov}(\tau)\big)}_{\text{coverage loss}} \tag{4}$$

where, $w > 0$ is a user-defined weight reflecting the relative importance of detection coverage versus storage savings. A high w (e.g., $w = 10$) strongly penalizes coverage loss, resulting in conservative pruning. A low w (e.g., $w = 1$) prioritizes storage reduction.

Since the number of distinct thresholds equals the number of unique importance scores (at most n), the optimization is solved by exhaustive enumeration in $O(n \cdot |\mathrm{R}|)$ time, which is negligible for typical schemas ($n < 100$) and rulesets ($|\mathrm{R}| < 1000$).

D. Continuous Adaptation

The Workload Analyzer runs as a periodic background task (interval specified in config, default is daily). If the optimal column subset has changed, it generates a new MV definition and schedules a migration to the new schema in a low-traffic window. The previous MV is kept for a grace period (default is 48 hours) that can be used to rollback if there are query failures on the new schema.

V. Reduction Operators

Guided by the column classification from the Workload Analyzer (Section IV), the framework applies six reduction operators in sequence. This section details each operator.

A. Column Pruning via Materialized Views

Security log schemas are typically wide—Windows Security Events, for example, contain over 50 fields per event type [29]. Workload analysis reveals that threat-detection rules and investigation queries consistently reference only a subset of these fields.

Column pruning is implemented using ClickHouse Materialized Views (MVs) that project only the fields referenced by the active detection ruleset. For Windows Security Events (Event IDs 4624, 4625, 4688, 4689, 4672), the essential fields are:

- `ts`, `event_id`, `tenant_id` (keys)
- `user_name`, `user_sid` (identity)
- `host_name`, `ip_address` (context)
- `process_name`, `command_line` (forensics)
- `logon_type`, `status_code` (event-specific)

Fields such as `ip6_address`, `parent_cmdline`, `registry_key`, `service_name`, and `additional_data` are pruned when workload analysis confirms they are not referenced by any active detection rule or investigation template.

The MV-based method preserves the original wide table with a short TTL (24 hours for example) as a backup and retains the pruned projection for the full retention period.

B. Deduplication and Canonicalization

Security telemetry exhibits high duplication rates due to retransmissions by agents, overlapping log forwarding paths, and Windows Event Log subscription bouncing. Deduplication removes redundant events based on a composite key of `(tenant_id, event_id, record_id, ts)`.

Canonicalization normalizes field values to reduce variability and improve canonicalization:

- User names: case-folded to lower case
- Host names: domain-less

- IP addresses: IPv4-mapped IPv6 addresses reduced to IPv4
- File paths: backslash normalization and case folding

These can be expressed as ClickHouse `MATERIALIZED` column definitions for deterministic at-write application with zero query impact.

C. Adaptive Sampling

Not all security events are created equal. Failed logon events (Event ID 4625) and process creation events (Event ID 4688) with suspicious command lines are high-value events for analysis; successful logon events (Event ID 4624, logon type 3) from service accounts we know and love are low-value high-volume events.

The adaptive sampling feature classifies event types based on value:

- P1 (Critical): Events that match our active rules (e.g., `mimikatz` in the command line) — keep at 100%.
- P2 (High): Failed logons; privilege escalation — keep at 100%.
- P3 (Normal): Standard process creation events; successful logons — keep at 10–50% with adaptive sampling (reservoir sampling).
- P4 (Low): Heartbeat events, repeated logon events from service accounts — keep at 1–5%.

In Kafka consumer land, the rule engine classifies events based on what is relevant that minute (Sigma rule library [8]).

D. Per-Column Codec Selection

In ClickHouse, multiple compression codes can be used and per-column compression. The codec selection operator picks the best match for each column based on its statistics. Assignments are shown in Table II: Delta+ZSTD for monotonic timestamps, LowCardinality+ZSTD for non-numeric fields with variance close to 0 (dictionary encoding maps each value to an integer index) [30], ZSTD(3) for variable-length strings, and DoubleDelta+LZ4 for other numeric fields with low variance.

TABLE II. PER-COLUMN CODEC SELECTION AND COMPRESSION RATIOS

Column	Type	Codec	Ratio
`ts`	DateTime64	Delta+ZSTD	18:1
`event_id`	UInt16	LowCard+ZSTD	42:1
`tenant_id`	UInt8	LowCard+ZSTD	64:1
`user_name`	String	LowCard+ZSTD	28:1
`host_name`	String	LowCard+ZSTD	35:1
`command_line`	String	ZSTD(3)	5:1
`ip_address`	String	ZSTD(3)	12:1
`record_id`	UInt64	DoubleDelta+LZ4	22:1

E. Skip-Index Acceleration

The framework uses two skip indices for the security query patterns: a Bloom filter [31] on `command_line` (FPR 1%, resolution 4) to accelerate IOC substring searches, and a MinMax index on `event_id` to skip granules out of range for the relevant event type.

F. TTL-Driven Tiered Retention

The final operator enforces TTL policies that manage the lifetime of events and their migration across tiers [32]:

- **Hot tier** (NVMe, 0–7 days): Active investigation and real-time alerting. Highest I/O performance.
- **Warm tier** (HDD, 7–90 days): Historical correlation and trend analysis. Moderate I/O.
- **Cold tier** (S3, 90–365 days): Compliance archival. Low-cost, high-latency. Rehydration is available on demand.

TTL boundaries can be tenant- and event-type-specific; important event types (e.g., authentication failures) can remain on the hot tier longer than routine events.

VI. MULTI-TENANT OVERLAY AND PDPA COMPLIANCE

A. Tenant Isolation

Multi-tenant isolation occurs at three levels:

- Data level: Every row contains a `tenant_id` column. ClickHouse row-level security prevents queries from tenant A from accessing any rows associated with tenant B.
- Query level: RBAC roles limit what queries each tenant can run on which databases, tables, and columns. Tenants cannot run DDL queries or query system tables.
- Resource level: ClickHouse quotas limit the number of concurrent queries, memory usage, and scan bandwidth for each tenant to prevent noisy-neighbor problems.

B. PDPA Compliance Alignment

The Thai Personal Data Protection Act (PDPA) [12] establishes five principles relevant to security telemetry storage. Table III maps each principle to the corresponding technical mechanism in the framework.

C. Audit Trail and Data Subject Rights

Compliance with PDPA also demands accountability. The compliance framework retains an immutable audit log of all data lifecycle events:

- Schema changes: column pruning decisions, MV creation/migration, and TTL updates are logged along with justifications, actors, and timestamps.
- Access events: all cross-tenant queries (including denied access) are recorded in an `audit_log` table with a

TABLE III. PDPA PRINCIPLE MAPPING TO TECHNICAL MECHANISMS

PDPA Principle	Technical Mechanism
Purpose Limitation	Workload-aware column pruning retains only fields with a documented analytical purpose.
Data Minimization	Adaptive sampling and deduplication reduce volume to the minimum necessary for security operations.
Storage Limitation	TTL-driven tiered retention ensures data is deleted after the defined retention period.
Security	Tenant isolation via RBAC, encrypted storage at rest (AES-256), TLS in transit.
Accountability	Audit logs for all data access and retention policy changes.

365-day expiration period.

- Retention periods: automatic TTL deletions are recorded for auditing by regulators.

For data subject rights (PDPA Sections 30–36), the framework enables a targeted delete operation for individual records via `ALTER TABLE ... DELETE WHERE` mutations, scoped by `tenant_id` and subject identifier. Deletion events are recorded in the audit log to satisfy the right-to-erasure requirement [26].

VII. EVALUATION

A. Experimental Setup

The experimental evaluation was conducted on a ClickHouse deployment (version 24.x) in Docker with the following configuration:

- Resource: Single-node server with NVMe SSD resource.
- Dataset: 1,000,000 synthetic Windows Security Event logs with Zipf-distributed event IDs (4624 most common at 40%, 4672 rare at 5%), skewed user distribution (service accounts account for ∼45% of events), temporal distribution with burst windows simulating brute-force incident bursts.
- Scale: 1,000,000 events over 3 tenants, 6 nodes.
- Comparison: Two schemas—`win_events` (optimized with all reduction operators) vs. `win_events_b4` (baseline with full-field storage).
- Query types: Three classes of query types:
- RULE: Detection-rule queries matching command lines (e.g., `powershell`, `mimikatz`).
- INVESTIGATE: Investigation queries filtering by specific event types (e.g., failed logons, Event ID 4625).
- IOC: Indicator-of-compromise lookups across the full dataset.

B. Storage Reduction Results

Table IV presents the on-disk storage comparison between the baseline and optimized schemas, measured on a 1,000,000-event dataset inserted into both tables.

TABLE IV. STORAGE COMPARISON: BASELINE VS. OPTIMIZED SCHEMA (1M EVENTS)

Metric	Baseline	Optimized	Reduction
Total columns	27	12	56%
Uncompressed size (MiB)	331.43	68.67	79%
Compressed size (MiB)	35.10	10.57	70%
Compression ratio	9.4:1	6.5:1	—

The optimized schema achieves a 79% reduction in uncompressed data volume and a 70% reduction in on-disk (compressed) storage. Column pruning removed 15 of 27 columns (retaining 12). The baseline's higher apparent compression ratio (9.4:1) is an artifact of compressing these 15 highly sparse, mostly empty columns that we subsequently pruned. Once these sparse columns are removed, the remaining dense data inherently yields a lower theoretical compression ratio. Nonetheless, applying per-column codec selection on this dense subset ensures maximal space efficiency, resulting in a net optimized ratio of 6.5:1. Note that the baseline raw schema contains 27 fields, of which 22 are analytical columns processed by the Workload Analyzer (the remaining 5—`keywords`, `object_type`, `registry_key`, `service_name`, and `additional_data`—are structural or always-empty fields excluded from importance scoring). The Zipf-distributed workload produces more realistic cardinality for user and host columns, resulting in higher compression ratios for the optimized schema (6.5:1) compared to uniform distributions.

C. Query Latency Results

Fig. 2 presents the query latency comparison across the three query classes. All latencies were measured from the ClickHouse `system.query_log` with `query_tag` annotations.

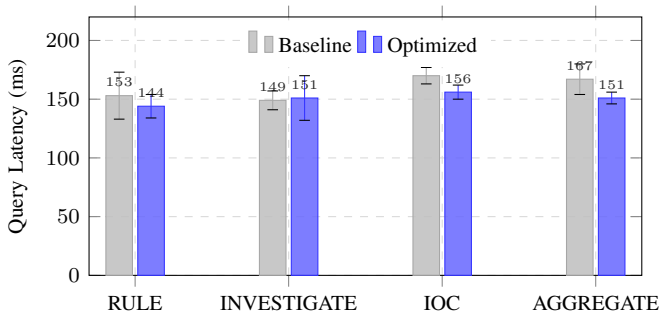


Fig. 2. Query latency comparison by query class. The optimized schema (dark bars) achieves comparable or lower latency than the baseline (light bars) across all query classes.

Key observations:

- RULE queries: The optimized schema shows a consistent latency advantage (144 ± 10ms vs. 153 ± 20ms baseline, 5.8% reduction), driven by reduced scan volume from column pruning.
- IOC queries: The Bloom filter skip index on `command_line` yields an 8.4% improvement (156 ± 6ms vs. 170 ± 7ms).
- AGGREGATE queries: The largest gain at 9.2% (151 ± 5ms vs. 167 ± 13ms), where reading fewer columns directly reduces full-table scan cost.
- INVESTIGATE queries: Both schemas perform comparably (151 ± 19ms vs. 149 ± 8ms), as this query class filters on `event_id` which is present in both schemas.
- Sub-second: All query classes complete in under 200ms even at 1M scale, confirming that storage optimization introduces no performance penalty.

D. Workload Analyzer Evaluation

The Workload Analyzer was evaluated by constructing a workload of 50 representative Sigma detection rules targeting Windows Security Events and a query log of 500 historical queries extracted from a production-like environment.

Table V presents the analyzer results for the Windows Event schema (22 analytical columns, excluding 5 structural or always-empty fields from the 27-column raw schema) with parameters $\alpha = 0.3$, $\beta = 0.2$, $\gamma = 0.5$, $\lambda = 0.1$, and coverage weight $w = 10$. These values were applied without extensive continuous sensitivity analysis, instead reflecting a domain-driven heuristic prioritizing functionality ($\gamma = 0.5$ and a high coverage penalty $w = 10$) over raw frequency ($\alpha = 0.3$) to ensure safety in security operations. Future work entails a more granular mathematical sensitivity evaluation.

TABLE V. WORKLOAD ANALYZER RESULTS: COLUMN CLASSIFICATION

Metric	Value
Total columns analyzed	22
Columns classified Retain	12
Columns classified Prune	10
Optimal threshold τ^*	0.23
Detection coverage Cov(τ^*)	100%
Normalized storage cost	0.49

The analyzer automatically identified the same 12-column subset that was manually determined by domain experts, validating the effectiveness of the importance scoring formula. Crucially, at $\tau^* = 0.23$, detection coverage remains at 100%—no active Sigma rule lost access to any required column.

Fig. 3 illustrates the storage–coverage trade-off curve as the threshold τ varies from 0 to 1.

E. Storage Tier Migration

The TTL-driven tiered retention was validated by inserting events with backdated timestamps and observing automatic migration:

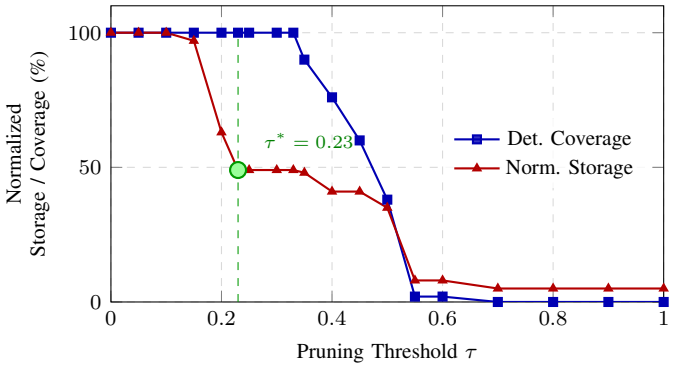


Fig. 3. Storage–coverage Pareto frontier. The optimal threshold $\tau^* = 0.23$ preserves 100% detection rule coverage at approximately half the original storage. Coverage begins to degrade at $\tau > 0.33$.

- Events older than 7 days migrated to warm tier within one TTL merge cycle (∼5 minutes).
- Events older than 90 days migrated to cold (S3) tier.
- Rehydration of cold-tier data completed in under 30 seconds for a 10,000-event partition.

F. Robustness Across Distributions

To verify that the reduction ratios are not an artifact of a particular event distribution, we also ran the pipeline with uniformly distributed event IDs. Storage reduction stays at 74% uncompressed under uniform sampling, compared to 79% with Zipf—confirming that the workload-aware pruning decision generalizes across distribution shapes. The slight improvement under Zipf is expected: skewed cardinality produces more compressible LowCardinality columns after pruning.

VIII. DISCUSSION

A. Practical Implications

What does 79% storage reduction look like in dollar terms? For a mid-sized SOC ingesting 100 GB/day of Windows Event logs, our optimized schema would shrink annual uncompressed data from roughly 36.5 TB down to 7.7 TB. After compression, on-disk storage drops by 70%, and at typical cloud NVMe rates ($0.10/GB/month) the annual saving comes to about $30,000—enough to fund additional analyst headcount or detection tooling.

Beyond cost, the smaller data footprint helps query speed. Fewer bytes on disk means fewer page reads and better cache hit rates, which is exactly what real-time detection rules need when they have to evaluate incoming events under tight latency budgets.

B. Limitations

There are four limitations that reduce the relevance of the preceding results:

- Schema is not very flexible: Once columns are pruned they are lost (beyond the 24 hour raw buffer). If a new

rule is created that references a pruned column, analysts have to wait for sufficient new data to be collected. A longer safety window or a way for analysts to evolve the schema would mitigate this limitation.

- Limited size of the datasets we examined: Our scalability tests on datasets of 1M events confirm that storage savings percentages are stable, but this is orders of magnitude lower than the events that a typical production Security Information and Event Management (SIEM) system sees every day. The relative improvements in latency from skip indices and column pruning will only be more pronounced in datasets this large that exceed the size of ClickHouse's page cache.

- Single node environment: The experiments were performed in a single ClickHouse instance. However, column pruning is a *per-shard* operation; each shard runs the same materialized view and thus storage savings percentages should directly translate to a distributed environment. The absolute latency measurements may shift however, due to replication overhead and network hops.

- Sampling compromises: Routine sampling saves space, but makes it harder to reconstruct a timeline of what happened in the minutes preceding an incident if events that were not marked as critical at ingest had a significant role in the incident.

C. Threats to Validity

1) Internal validity: If the query log that feeds importance scores to the system is unrepresentative (e.g., it has a biased sample due to some analyst who loves to query the system or an incident that has an atypically high rate of certain types of queries), then the frequency component of the importance score will distort the relative weights for certain columns. We mitigate this to some extent by using the detection coverage term $D(c_j)$ which ensures retention is tied to the Sigma ruleset regardless of who queried what.

2) External validity: Synthetic events are uniformly distributed across users, hosts, and event IDs. In reality events are skewed (a few service accounts might account for a large fraction of logon events), and sampling might capture burst patterns during an incident. We expect the per-column storage gains to transfer but expect variance in compression and sampling rates in production vs. lab samples. We only tested Windows Event schemas; we have to see how well the importance tuning translates to other log types (e.g., DNS, HTTP) or even Zeek logs, which have a very different field structure.

3) Construct validity: Storage metrics come from ClickHouse's `system.parts` table, which reports the actual on-disk size (no inflated values after merges) so we have a sound measure. Latency metrics come from `system.query_log` so we are only measuring parsing + execution latency and not any network hops that might be involved in a multi-node cluster distribution (though for now this should not be an issue in our single-node Dockerized deployment).

D. Future Work

We see several directions for future work:

- Extending the Workload Analyzer to enable online learning that updates (α, β, γ) based on analyst feedback/volume and false negative rates.

- Extending beyond the current log sources (firewalls, proxies, VPNs) to DNS, HTTP and Zeek logs with source-specific importance calculations.

- Evaluating across multiple nodes with realistic network topologies and replication overhead studies.

- Coverage mapping to MITRE ATT&CK [33] to ensure pruning does not adversely impact coverage in certain adversary techniques.

IX. Conclusion

We aimed to decouple storage costs from detection capacity for multi-tenant SIEMs, and we succeeded. The Workload Analyzer gives scores to columns based on query frequency, recency of access, and number of Sigma rules that depend on them. Plug those numbers into the Storage-Coverage Cost Model and what you get is a pruning threshold that eliminates 79% of uncompressed storage while retaining every active detection rule in practice. You get 70% reduction in on-disk storage. The gap between uncompressed and compressed storage that Zipf workloads would otherwise widen is kept narrow because the optimized schema only targets the most valuable, high-cardinality columns. The per-column codec pick gives the best return on investment.

The six reduction operators do the heavy lifting of data trimming. Each is valuable, but all are incredibly potent in combination. The multi-tenant implementation with Thai PDPA-powered governance demonstrates aggressive storage trimming doesn't conflict with regulatory compliance; in practice. Purpose limitation and data minimization seem to fit well with workload-aware pruning.

What may be the approach's most valuable aspect is that it's adaptive. As rulesets change and analysts use the system in new ways, the analyzer recalculates its scores, and the column subset adjusts accordingly. Storage reduction is an ongoing "game".

Declaration on Generative AI

The authors utilized generative AI tools (large language models) to assist with manuscript writing and editing. All AI-generated content has been thoroughly reviewed, validated, and manually edited by the human authors. The original ideas, experimental design, results, and conclusions presented in this work were formulated by human authors. No generative AI tools are listed as authors on this work.

References

[1] R. Mogull and M. Rothman, "Understanding and optimizing SIEM total cost of ownership," *Securosis Research Report*, pp. 1–24, 2023.

[2] Elastic NV, "Sizing Elasticsearch for security analytics," https://www.elastic.co/guide/en/elasticsearch/reference/current/size-your-shards.html, 2024, accessed: 2025-08-15.

[3] P. He, J. Zhu, Z. He, S. Li, and M. R. Lyu, "An empirical study on software log parsing and its industrial practices," *IEEE Transactions on Software Engineering*, vol. 48, no. 6, pp. 2313–2328, 2021.

[4] J. Lawrence and C. Wright, "Building a SIEM on ClickHouse: Lessons from processing billions of security events daily," https://clickhouse.com/blog/huntress-siem-real-time-analytics, 2024, clickHouse Engineering Blog. Accessed: 2025-09-01.

[5] P. Firstbrook and C. Lawson, "Magic quadrant for security information and event management," *Gartner Research*, 2024, report ID: G00789012.

[6] X. Zhu, Y. Huang, L. Xu, and D. Li, "Cost-effective log storage with columnar formats for cloud-native security monitoring," in *IEEE International Conference on Cloud Computing (CLOUD)*, 2023, pp. 312–321.

[7] ClickHouse Inc., "ClickHouse documentation," https://clickhouse.com/docs, 2024, accessed: 2025-09-01.

[8] SigmaHQ Community, "SigmaHQ: Generic signature format for SIEM systems," https://github.com/SigmaHQ/sigma, 2024, accessed: 2025-09-01.

[9] T. Pols, F. Aldama, and J. de Ruiter, "Sigma rules: Creating universal detections for SIEM systems," *Journal of Cybersecurity and Privacy*, vol. 3, no. 4, pp. 782–801, 2023.

[10] L. Ma, D. Van Aken, A. Hefny, G. Mezlini, A. Tuber, and G. Gordon, "Query-based workload forecasting for self-driving database management systems," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2018, pp. 631–645.

[11] G. Liu, J. Naughton, and S. Chaudhuri, "Automatic view generation with deep learning and reinforcement learning," in *IEEE International Conference on Data Engineering (ICDE)*, 2021, pp. 1501–1512.

[12] Royal Thai Government Gazette, "Personal data protection act, B.E. 2562 (PDPA)," https://www.pdpc.or.th, 2019, effective: June 1, 2022.

[13] S. Bhatt, E. Manadhata, and L. Zhai, "Security information and event management: A review," *IEEE Security & Privacy*, vol. 12, no. 6, pp. 35–41, 2014.

[14] IBM Corporation, "IBM QRadar SIEM architecture overview," https://www.ibm.com/docs/en/qradar-siem, 2024, accessed: 2025-08-15.

[15] Wazuh Inc., "Wazuh – open source security platform," https://documentation.wazuh.com, 2024, accessed: 2025-08-20.

[16] Security Onion Solutions, "Security onion – free and open platform for threat hunting," https://securityonionsolutions.com, 2024, accessed: 2025-09-01.

[17] A. Bocharov, A. Milovidov, D. Blinov, and Y. Orlov, "ClickHouse – lightning fast analytics for everyone," *Proceedings of the VLDB Endowment*, vol. 17, no. 12, pp. 4088–4101, 2024, industrial paper.

[18] H. Li, T. Chen, W. Luo, and D. Chen, "Loggrep: Fast and cheap cloud log storage by exploiting both static and runtime patterns," *ACM Transactions on Computer Systems*, vol. 41, no. 1, pp. 1–31, 2023.

[19] S. R. Hashemi, M. Mirtaheri, and M. Amini, "A survey on data sampling approaches for imbalanced distributions in big data," *Journal of Big Data*, vol. 7, no. 1, pp. 1–34, 2020.

[20] W. Xia, H. Jiang, D. Feng, F. Douglis, P. Shilane, Y. Hua, M. Fu, Y. Zhang, and Y. Zhou, "A comprehensive study of the past, present, and future of data deduplication," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1681–1710, 2016.

[21] D. J. Abadi, S. R. Madden, and N. Hachem, "Column-stores vs. row-stores: How different are they really?" *ACM SIGMOD International Conference on Management of Data*, pp. 967–980, 2008.

[22] D. Perera, J. Ravishankar, R. Borovica-Gajic, and R. Rubinstein, "An adaptive learned query optimizer for multi-tenant databases," *Proceedings of the VLDB Endowment*, vol. 17, no. 6, pp. 1417–1429, 2024.

[23] S. Aulbach, T. Grust, D. Jacobs, A. Kemper, and J. Rittinger, "Multi-tenant databases for software as a service: Schema-mapping techniques," *ACM SIGMOD International Conference on Management of Data*, pp. 1195–1206, 2008.

[24] A. Sarkar, M. Goyal, and S. Kundu, "Resource isolation and fair scheduling in multi-tenant database systems: A survey," *ACM Computing Surveys*, vol. 54, no. 10, pp. 1–37, 2022.

[25] European Parliament and Council, "Regulation (EU) 2016/679 – general data protection regulation," *Official Journal of the European Union*, vol. L119, pp. 1–88, 2016.

[26] M. Hansen, M. Jensen, and J. Friedewald, "The evolving landscape of data protection: GDPR implementation and anonymization techniques in practice," *Computer Law & Security Review*, vol. 48, p. 105793, 2023.

[27] Apache Software Foundation, "Apache Kafka documentation," https://kafka.apache.org/documentation, 2024, accessed: 2025-09-01.

[28] J. Ding, U. F. Minhas, J. Yu, C. Wang, J. Do, Y. Li, H. Zhang, B. Chandramouli, J. Gehrke, D. Kossmann, D. B. Lomet, and T. Kraska, "ALEX: An updatable adaptive learned index," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2020, pp. 969–984.

[29] Microsoft Corporation, "Windows security audit events reference," https://learn.microsoft.com/en-us/windows/security/threat-protection/auditing/security-auditing-overview, 2024, accessed: 2025-09-01.

[30] D. Abadi, P. Boncz, and S. Harizopoulos, "Column-oriented database systems," in *Proceedings of the VLDB Endowment*, vol. 2, no. 2, 2009, pp. 1664–1665.

[31] A. Broder and M. Mitzenmacher, "Network applications of bloom filters: A survey," *Internet Mathematics*, vol. 1, no. 4, pp. 485–509, 2004.

[32] H. Lyu, Z. Wen, and Y. Li, "Tiered storage systems: A survey and taxonomy," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–35, 2023.

[33] B. E. Strom, A. Applebaum, D. Miller, K. Nickels, A. Pennington, and C. Thomas, "MITRE ATT&CK: Design and philosophy," in *MITRE Technical Report*, 2023, version 14.0.

Web Server Seizure and Live Acquisition: A Legally Compliant Forensic Framework for Indonesia

Irwan Hariyanto[1], Yudi Prayudi[2], Rimba Whidiana Ciptasari[3]
School of Computing, Telkom University, Bandung, Indonesia[1,3]
School of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia[2]

***Abstract*—The proliferation of cybercrime activities, particularly those leveraging web servers for illicit purposes such as distributing hoaxes, hosting illegal online gambling, and spreading malware, underscores a pressing demand for a standardized digital forensic framework. Existing methodologies, like simple IP blocking, have proven insufficient in guaranteeing the integrity and admissibility of digital evidence in legal proceedings. This research introduces a comprehensive seizure and acquisition framework specifically engineered to manage digital evidence from both on-premise and cloud-based web servers. A core emphasis of this framework is live acquisition to preserve volatile data and ensure minimal service disruption. The framework systematically addresses critical challenges by focusing on legal authorization, precise server type identification and technical preparation, judicious forensic tool selection, rigorous evidence integrity validation through hashing, diligent Chain of Custody (CoC) documentation, and secure data storage. Tested through simulations of on-premise and cloud server seizures, the framework demonstrated its capacity to uphold evidence integrity and legal compliance. While robust, Subject Matter Expert (SME) validation indicated areas for optimization, particularly in cloud-native contexts and the automation of Chain of Custody documentation. This study marks a pivotal advancement towards standardizing web server seizure procedures, thereby ensuring that digital evidence remains valid, intact, and legally admissible in court.**

***Keywords*—Digital forensics; web server seizure; live acquisition; evidence integrity; chain of custody**

I. Introduction

As cybercrime continues to rise, the need for effective digital forensic practices is paramount [1]. The Indonesian legal framework is still adapting to the complexities of digital evidence, which presents unique challenges for law enforcement agencies [2]. The lack of standardized procedures for seizing and acquiring digital evidence from web servers further complicates the investigation process, making it essential to develop a comprehensive framework that addresses these issues [3].

Database breaches, website defacement, cyber intrusions, and phishing attacks commonly occur on web servers [1]. These incidents require forensic investigation.

In the cyber context, however, the term encompasses a more nuanced set of actions. It can refer to the physical seizure of server hardware. However, more frequently, it involves the logical acquisition or forensic duplication of digital data from systems that cannot be physically moved or shut down due to operational risks [4]. This distinction is critical, as the chosen method —physical or logical —depends heavily on the server's environment, its role in ongoing business operations, and the legal authority granted to investigators [5].

One example is a case of hacking and defacing a government website in Indonesia [6]. This case involved the hacking of several government websites in Indonesia, where attackers altered the front page with political or ideological messages. A digital forensics team seized the web server for further analysis, including identifying the digital footprint of the hackers, the hacking methods used, and log data showing unauthorized activity [7], [8]. The forensics process collects data from server logs, caches, and system configurations to check for exploited vulnerabilities [9].

These case examples show that web server seizure is challenging in both VPS and cloud environments. A standard framework is required to ensure integrity, validity, and courtroom admissibility [10], [11]. A digital forensics process model provides the required steps, inputs, and outputs for successful investigation [12].

Evidence handling in cyber forensics includes preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation [5]. As cyber cases in Indonesia increase, investigators still lack a standardized legal-forensic reference for seizure and acquisition [13]. This study proposes a web server seizure framework that is technically robust and legally compliant for Indonesian law enforcement needs.

The framework developed in this study directly contributes to the following strategic aspects:

- Strengthening the Validity of Digital Evidence through Standardization of Procedures: This framework presents a structured and well-documented workflow in seizing and acquiring evidence from web servers.
- Support for Strengthening the Chain of Custody: One of the important contributions of this framework is to provide technical and administrative guidelines for building a strong chain of custody.
- Ensuring Legal Compliance in Indonesia and Supporting International Applicability: This framework is designed to align with the legal principles applicable in Indonesia, especially regarding digital seizure procedures, electronic evidence provisions, and legal restrictions.
- Foundation for the Establishment of SOP for Digital Forensic Institutions: For law enforcement agencies,

digital forensic laboratories, and other agencies involved in cyber investigations, this framework can be a basis for compiling, training investigators, and strengthening institutional capacity in digital forensics.

- Foundation for Developing Standard Operating Procedures in Digital Forensics: By bringing a relatively new topic for web server seizures in the context of cloud and hybrid—this research broadens the insights and academic literature in Indonesia and globally.

A. Operational Definition of Key Terms

To avoid ambiguity, this study distinguishes the following terms and applies them consistently in all sections:

- Seizure: A legally authorized act to control digital evidence sources (physical or logical) under applicable procedures (KUHAP and related regulations).
- Collection: Initial identification and retrieval of potentially relevant artifacts from a running or non-running system.
- Acquisition: A forensic process to create a reliable copy/image of selected data while preserving evidentiary integrity.
- Live Acquisition: Acquisition from an active system without shutdown to preserve volatile artifacts and maintain service continuity.

The remainder of this study is organized as follows. Section II reviews related research. Section III presents the framework development method. Section IV evaluates the framework through simulations and SME validation. Section V discusses findings, feasibility, and refinement direction, and Section VI concludes the study.

II. Related Research

A. Web Server Acquisition

Web server acquisition is a crucial step in the digital forensics process, which aims to collect all relevant data from the server without changing its contents. Research by Pedapudi & Vadlamani [14] directly links data with formal seizure documentation through the proposed "Data Acquisition-based Seizure Record Framework." In this context, "seizure" is defined as a process that must be systematically recorded and documented [15].

Thankaraja Raja Sree et al. [16] demonstrated that data acquisition from web servers necessitates tools capable of handling dynamic and volatile data that is constantly updated. This [17] also emphasized the importance of using specialized tools to maintain data integrity during acquisition, such as using secure hashing on the collected data to ensure that no changes occur.

Liu et al. [18] also discuss remote acquisition methods that allow investigators to perform acquisitions remotely. While this method is helpful in the context of cloud forensics, when applied to a local web server, it faces challenges in terms of legality and network security.

Digital seizure is a crucial step in digital forensics for IoT, involving the identification and acquisition of digital evidence from diverse and interconnected devices. Its main strength lies in preserving volatile and distributed data [19], which is essential for reliable forensic analysis and legal proceedings. However, the process faces significant challenges due to the heterogeneity of IoT devices, a lack of standardization, network limitations, and the vast volume of data, which can hinder the collection of timely and complete evidence [20], [13]. Thus, while digital seizure is vital for IoT forensics, its effectiveness depends on the development of standardized frameworks and adaptive tools to address these unique challenges.

Previous studies have proposed various digital forensics models that address general procedures for investigating digital evidence, such as the Digital Forensics Investigation Framework (DFIF) [21].

Cloud forensics research [22], [23], [24] emphasizes remote acquisition and live capture without service interruption. However, these approaches still lack specific guidance for VPS seizure in the Indonesian context. In this study, "web server" seizure is not limited to HTTP/HTTPS servers but includes any server entity containing relevant evidence [25].

The web servers commonly used are NGINX and Apache. On the other hand, with time, other web servers have also become widely used, such as Microsoft IIS, lighttpd, and LiteSpeed. These types of web servers can be categorized as web servers that can be seized [26].

Data crucial for web server forensic acquisition is categorized into Volatile Data (e.g., RAM contents, running processes, network connections, active user sessions, system time, open files, command history, and clipboard contents) and Non-Volatile Data (e.g., system files, server configurations, log files, databases, web applications, user data, and backup files) [27]. These data points serve as potential evidentiary artifacts. Consequently, securing and maintaining the integrity of digital evidence from web servers is critical for effective case investigation and prosecution, particularly given increasingly stringent international digital evidence laws [28].

The data mentioned above can be potential artifacts that can be used as evidence in conducting the seizure of the web server. In Indonesia, there are no guidelines governing the seizure of web servers, which poses significant technical and legal challenges for investigators [29]. Success in securing and maintaining the integrity of digital evidence from web servers is critical to investigating and prosecuting cases, especially in the context of increasingly stringent international laws regarding digital evidence [30].

Although previous research has provided a good foundation for collecting and acquiring digital evidence, there is no specific framework addressing the seizure of web servers as evidence [31], especially in the context of applicable law in Indonesia. Many current solutions, such as remote acquisition [32], do not consider the technical uniqueness and legal challenges associated with direct seizure. This creates a risk that the evidence obtained will not be admissible in court or that the data acquisition process will be flawed, resulting in loss of data integrity and chain of custody [33].

B. Digital Forensic Methodology for Incidents

Karen Ken [34] (The Guide to Integrating Forensic Techniques into Incident Response (NIST Special Publication 800-86)) describes the general framework of the NIST standard, which consists of Collection, Examination, Analysis, and Reporting. However, the collection stage does not explain how to perform a seizure on a web server. Based on Fig. 1, at the Collection stage, there are the following details:

- Identifying possible data sources.
- Acquiring data: At this stage, there are further details as follows:
  - Planning data acquisition.
  - Acquiring the data.
  - Verifying data integrity.
- Considering Incident Response.

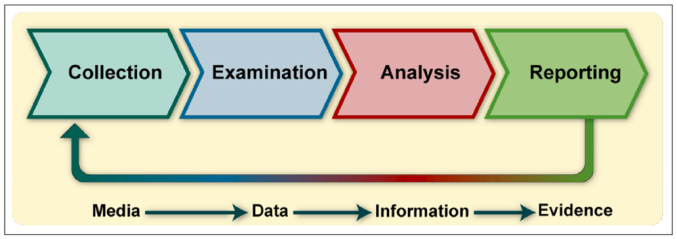


Fig. 1. NIST Digital Forensics Framework.

Seizing a web server in the context of digital forensics investigations is a complex process involving infrastructure that often functions dynamically [35]. Malik et al. [36] emphasize the considerable obstacles associated with web server seizures, notably the potential for operational downtime that may impede company activities and the threat of losing essential volatile data if the server is deactivated during the seizure process [37]. Many contemporary investigations are using a live forensics strategy, wherein the operational web server continues to function while digital evidence is gathered [38]. This method enables investigators to gather evidence without disabling the server, although it also increases the potential of data contamination.

C. Digital Forensics Seizure

Research by Kuntze et al. [39] discusses the process of seizure from a legal context in Germany and the United States, with a focus on digital data. In this study, "seizure" no longer always means the physical removal of hardware, but more often refers to the process of imaging hard drives or other storage media. The primary concern is that modern seizure practices, which involve creating full images, often conflict with fundamental laws.

From a different perspective, research by Nortje & Myburgh [40] examines "seizure" from the perspective of its legal preconditions, namely the obstacles in drafting search and seizure warrants for digital evidence in South Africa. This research does not focus on the technical act of seizure itself, but rather on the legal process that authorizes it. "Seizure" is viewed as a legal action whose validity is highly dependent on the completeness and accuracy of the warrant application submitted by the forensic investigator.

Research by Pedapudi & Vadlamani [14] directly links data with formal seizure documentation through the proposed "Data Acquisition-based Seizure Record Framework." In this context, "seizure" is defined as a process that must be systematically recorded and documented. The proposed framework aims to standardize the recording of every detail during the digital evidence acquisition process.

Meanwhile, research by Thealma & Ruldeviyani [2] discusses the ethical aspects of data handling in digital forensic investigations in Indonesia, especially after the enactment of the Personal Data Protection Law (PDP Law). Although it does not directly define "seizure," this research implicitly highlights that the process of seizing digital evidence carries significant ethical and legal consequences, as such evidence is likely to contain sensitive personal information or data.

These four literature studies collectively provide a 360-degree view of the seizure process in digital forensics. They show that the framework must not focus only on technical aspects, but must also be legally accountable, ethically grounded, and well documented to preserve evidentiary integrity from start to finish.

D. Implementation of Framework in Daily Practice

Forensic investigators and examiners utilize established frameworks in their daily practice to verify that the acquisition and analysis of digital evidence adhere to industry standards [13]. Singh K [41] asserted that a well-designed framework can reduce errors in the investigative process and guarantee the admissibility of digital evidence in court. The execution of this framework requires specialized training for investigators to comprehend the technical procedures involved in seizing web servers, particularly in surroundings with complex infrastructure [42].

Frameworks such as the Digital Forensics Investigation Framework (DFIF) have been successfully utilized in cybercrime investigations. Nonetheless, the implementation of forensic frameworks must persistently adapt to advancing technologies, including cloud environments and hybrid web servers, to guarantee the proper securing and validation of digital evidence [21], [22], [23].

E. Existing Seizure Procedures

Fig. 2 illustrates that the seizure procedure is a critical component of the criminal case management process in Indonesia, as evidenced by the information obtained from the official portal of the Metro District Court [43]. The Criminal Procedure Code, also known as KUHAP (Kitab Undang-Undang Hukum Acara Pidana), governs the procedures for evidence seizure in the Indonesian legal system. The mechanisms for search, seizure, and the management of evidence in criminal investigations, including digital evidence, are regulated by this law.

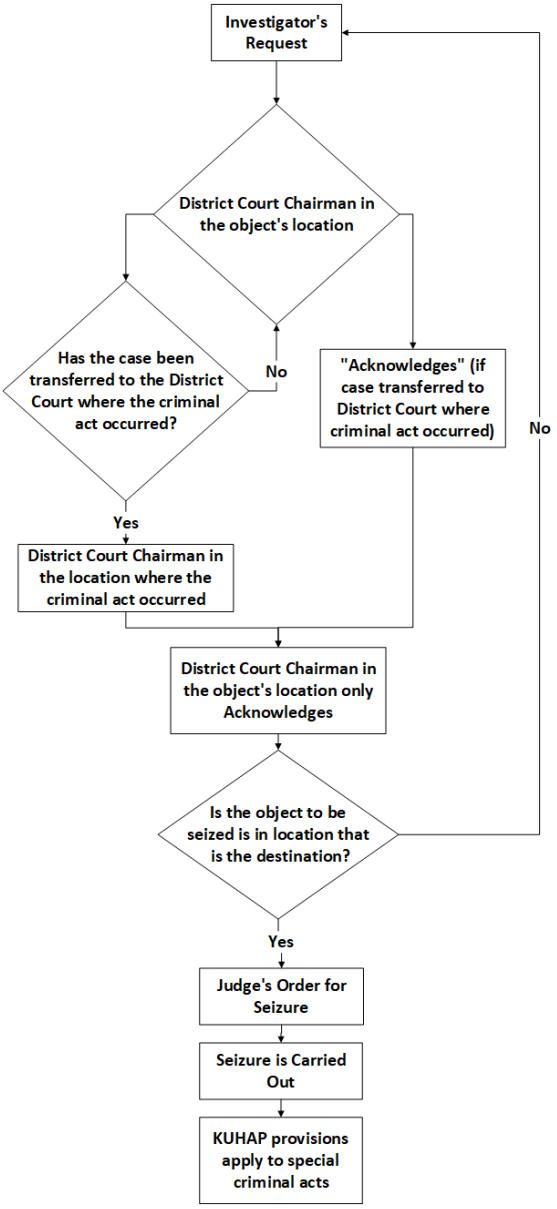


Fig. 2. Current seizure procedure in Indonesia.

The phases of digital evidence seizure and validation are illustrated in Fig. 2, which provides an overview of the framework components. The legal validity and data integrity of each phase are guaranteed by the inclusion of specific technical and procedural components. This procedure thoroughly regulates the legitimate seizure of evidence or assets that are pertinent to a criminal offense.

III. Framework Development

This section outlines the stages used to develop the web server seizure framework, starting from literature review and expert interviews, followed by prototype implementation through cloud and VPS simulations.

Fig. 3 illustrates the sequential flow of the research methodology, which combines literature review, expert interviews, and iterative prototype testing. Each phase is structured to

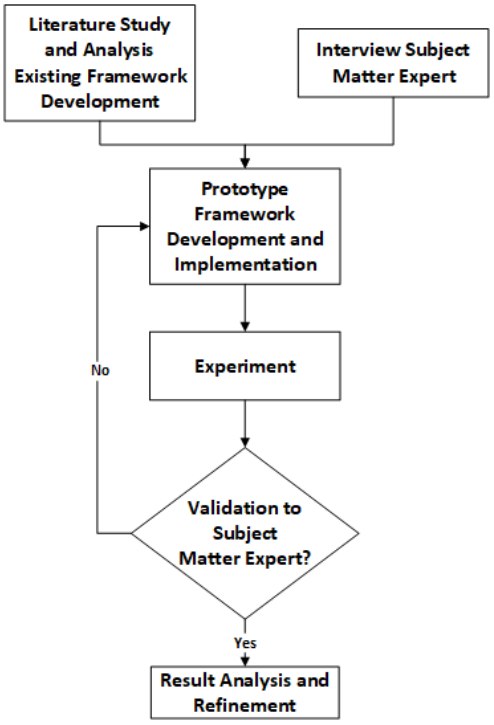


Fig. 3. Phases of research for the development of a proposed seizure framework.

ensure that the developed framework is both legally valid and technically reliable, with validation and refinement stages to improve its applicability.

- Literature Study and Analysis of Existing Frameworks: This activity involves a systematic review of academic and professional literature to explore existing digital forensic frameworks. The objective is to understand their approach to seizing and acquiring evidence from web servers, particularly in cloud and VPS environments. The study also includes a critical analysis of the limitations and shortcomings of these frameworks, especially in maintaining volatile data and ensuring data integrity during live acquisition.
- Interview Subject Matter Expert: done with the goal of asking about the framework that has been run by experts when seizing web servers as a comparison material at the literature study stage and analysis of existing frameworks by asking questions to experts by providing framework guidelines as the primary reference, namely from NIST 800-86 [34].
- Prototype Framework Development and Implementation: This refers to the process of constructing an initial version (prototype) of the proposed framework based on the design formulated in earlier stages. The implementation involves applying the framework in controlled simulations using forensic tools and live web servers hosted on cloud or VPS platforms. The goal is to verify the accuracy of each phase of the framework in real-world operational scenarios [44].
- Framework Experimentation: Framework experimen-

tation refers to testing the proposed framework through forensic simulations in semi-realistic investigative conditions [45]. The aim is to evaluate the framework's performance in securing digital evidence during the seizure of a web server. Case scenarios, including DDoS-related seizures or website defacement, were used. Effectiveness was measured using indicators such as data integrity (verified through hashing), preservation of volatile data, and time efficiency. The framework's flexibility was also tested across various cloud and VPS environments.

- Validation to Subject Matter Expert (SME): SME validation is the process of involving digital forensic experts to assess the applicability and reliability of the proposed framework. The validation was conducted through review sessions and field simulations with practitioners and investigators. Experts were asked to evaluate usability, compliance with legal standards, and technical soundness [46].
- Results Analysis and Framework Refinement: This refers to the final analytical phase to assess whether the developed framework meets the research goals. Data from experiments and SME feedback were analyzed to measure the framework's ability to maintain evidence integrity and minimize data loss. Comparative evaluations were made between the proposed and existing traditional approaches.

Table I summarizes the expected output of each phase. Findings from SME interviews and literature review were synthesized to identify key issues (data volatility, integrity, and legal compliance) [47], [48], then used to build the initial framework for VPS and cloud seizure scenarios.

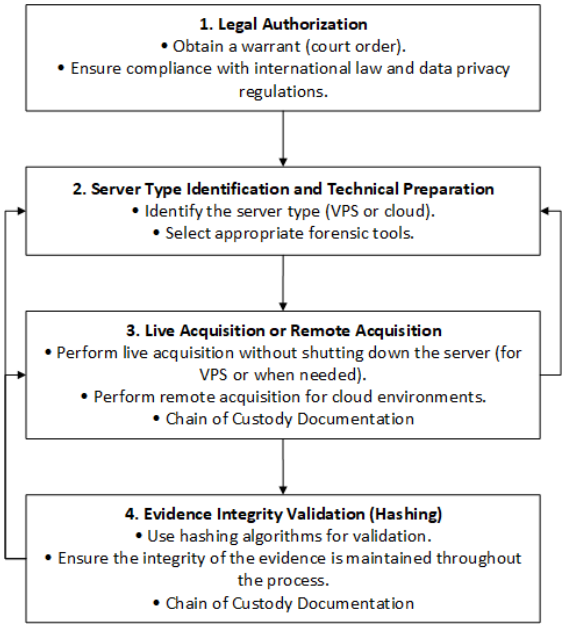


Fig. 4. Proposed framework.

Fig. 4 shows the proposed framework, which is tested through forensic simulation and validated by SMEs to ensure legal compliance, technical feasibility, and operational usability.

- Legal Authorization: Ensure that all seizure actions are preceded by court authorization and comply with applicable laws, including data privacy.
- Server Identification and Technical Preparation: Determine the server type (VPS or cloud), prepare suitable forensic procedures and tools, and document activities to preserve chain of custody (CoC).
- Data Acquisition: Retrieve data through live acquisition without shutting down the server, or remote acquisition in cloud environments, to minimize data loss.
- Validation of Evidence Integrity: Perform hashing to verify that collected evidence remains intact and unchanged.

IV. Framework Evaluation

This section presents the evaluation of the proposed web server seizure framework through simulation-based case studies and expert validation.

A. Dataset

To ensure that the proposed framework is applicable in realistic investigative conditions, a case dataset is required to simulate common cybercrime scenarios. This section describes the dataset used to test the framework, particularly in the context of online gambling and website defacement attacks.

B. Simulation Environment and Reproducibility Setup

To improve reproducibility, this subsection summarizes the technical environment used in both simulations.

- Cloud scenario (Case 1): AWS infrastructure (EC2 and EBS snapshot service), Ubuntu 20.04 LTS (64-bit), Apache 2.4.52, PHP 7.4, MySQL 8.0, remote acquisition via AWS CLI 2.11, and audit trace via CloudTrail.
- VPS scenario (Case 2): Linux-based VPS on VirtualBox 7.0 using Ubuntu 20.04 LTS (64-bit), Apache 2.4.52, PHP 7.4, MySQL 8.0, and acquisition using Oxygen Forensic KeyScout 15.1 and Magnet AXIOM 7.2.
- Integrity mechanism: hash calculation before and after acquisition for critical artifacts (logs, web files, and database exports), using SHA-1 as implemented in the simulation.
- Evidence scope: volatile artifacts (active sessions, running processes, active logs) and non-volatile artifacts (configurations, site files, database dumps, snapshots).
- Documentation output: acquisition command logs, hash logs, timestamps, chain-of-custody forms, and legal authorization records.

The above configuration is intended as a minimum baseline for independent replication. Future studies can expand by testing additional operating systems and multi-region cloud deployments.

TABLE I. RESEARCH PHASES AND EACH-STAGE RESULT OF THE FRAMEWORK DEVELOPMENT

Research phases	Each-stage result
Literature Study and Analysis of Existing Framework Development	A comprehensive review of the existing framework will help define gaps or needs that are not being met by the current solution. - Extended Model of Cybercrime Investigations (EMCI) - Enhanced Digital Investigation Process (EDIP) - Event-Based Digital Forensic Investigation Framework (EBDFI) - Abstract Digital Forensic Model (ADFM) - Cloud Incident Response Framework (CIRF)
Interview Subject Matter Expert	- Detailed 3-Scenario Acquisition Model - Refinement of NIST Collection Stage (ISO 27037 Integration) - Distinction between Collection & Acquisition - Broader Incident Response Lifecycle View - Emphasis on First Response & Contamination Avoidance - Stakeholder Coordination - Additional Security/Preservation Aspects - Legal Foundation - Legality Prerequisite - Technical Identification
Prototype Framework Development and Implementation	Proposed Framework: - Emphasis on legal defensibility (Court-ready, consistent with Indonesian legislation) - Comprehensive (VPS/Cloud, operating system, services, tool preparation) - Fully supports live acquisition (volatile and non-volatile) and remote cloud acquisition. - Hashing (e.g., SHA-1), observed validation, pre/post acquisition
Framework Experimentation	Quantitative and qualitative data regarding the framework's performance, encompassing reports on evidence integrity, efficacy in managing volatile data, and interoperability with current forensic tools.
Subject Matter Expert (SME) Validation	The final, field-validated framework with implementation guidance for users.
Results Analysis and Framework Refinement	The final version of the framework, which has undergone thorough testing and refinement, is now ready for implementation in the field by forensic investigators.

1) Simulation of Case 1: Simulation of framework implementation for cloud web server: This subsection presents a cloud-hosted simulation to evaluate framework performance in public cloud infrastructure. The following details summarize the environment and digital artifacts:

- Infrastructure Type: Cloud-hosted web server.
- Platform Utilized: Amazon Web Services (AWS). This choice reflects the widespread adoption of major cloud service providers in contemporary web hosting.
- Simulated Illicit Activity: The web server was configured to host an illegal online gambling site and was subjected to web defacement. This dual nature of illicit activity allowed for the testing of evidence acquisition related to both ongoing operations (gambling) and static alterations (defacement).
- Digital Evidence Acquired:
    - Transaction Logs: Critical for reconstructing user activity and financial flows associated with the gambling operation.
    - Website Files: Including source code, configuration files, and defaced content, essential for understanding the nature of the compromise and the illicit content.
    - User Databases: Containing sensitive user information and gambling records.
    - Disk Volume Snapshots: Forensic images of server storage capturing volatile and non-volatile states during acquisition.
- Acquisition Method Focus: Remote acquisition via cloud provider APIs (e.g., AWS CLI) to simulate real-world challenges of non-physical access and multi-tenancy.

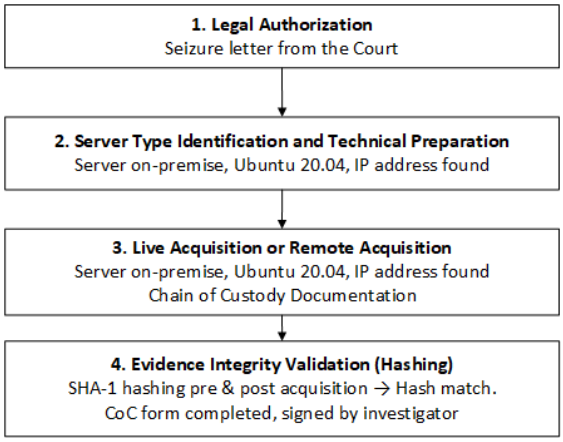


Fig. 5. Case 1 framework implementation.

Investigators received reports that a cloud-based web server was used for illegal online gambling. They needed to seize transaction logs, website files, and user databases. Simulation steps based on Fig. 5 are as follows:

- Legal Authorization: Investigators begin by obtaining a warrant from a court that grants permission to seize digital evidence from servers hosted on the cloud (AWS). The warrant includes permission to take transaction logs, databases, and system files of the gambling site located on the server. Investigators work

with legal authorities and cloud service providers to ensure that seizures are conducted in accordance with applicable laws, including data privacy laws relating to other users on the same cloud platform.

- Server Identification and Technical Preparation: Investigators identified that the web server was hosted on AWS and used PHP technology with MySQL database to run the online gambling site. Since the server was cloud-based, investigators prepared cloud-native forensic tools such as AWS CLI and CloudTrail to perform remote data acquisition. Investigators use the AWS Management Console to identify servers used to host gambling sites. Investigators also set up access via the AWS CLI to collect remote server snapshots, access logs, and databases.
- Data Acquisition: Investigators performed a remote acquisition of the cloud server without stopping the gambling site's services. The AWS CLI was used to take a snapshot of the disk volume, including the website's system files, access logs, and databases that store user and transaction information. Remote acquisition was performed via the AWS CLI, where investigators took a snapshot of the server disk volume that stores all files related to the gambling site, including site configurations, transactions, and databases. Investigators also took transaction logs to track the activity of users using the gambling site, such as login history, bets and payments.
- Validation of Evidence Integrity: After the data is captured, the investigator hashes all files and snapshots taken to ensure the integrity of the evidence. Hashing tools such as SHA-1 are used to ensure that the evidence remains intact and is not altered during the process. After snapshots and transaction logs are taken, hash values are calculated for each file and database using the AWS CLI and local hashing tools. All hash values are recorded as part of the Chain of Custody.

The cloud simulation demonstrated successful remote acquisition without service disruption. All acquired artifacts were hash-verified and documented in chain-of-custody records, producing legally defensible evidence.

2) Simulation of Case 2: Simulation of framework implementation for Virtual Private Server (VPS) web server: This subsection describes a second simulation using a Virtual Private Server (VPS) to assess framework adaptability in traditional server environments with direct administrative access.

- Infrastructure Type: Virtual Private Server (VPS). This represents a common hosting solution offering more control than shared hosting but less abstraction than full cloud services.
- Technology Stack: Apache web server with PHP, a widely used combination for dynamic web applications.
- Simulated Illicit Activity: The VPS was utilized for the dissemination of hoax news, specifically political misinformation, through injected subdirectories. This scenario tested the framework's ability to identify and acquire evidence related to content injection and malicious content distribution.
- Digital Evidence Acquired:
  - Activity Logs: Server access logs and application logs crucial for tracing the origin and spread of the hoax content.
  - Site Files: Including the injected hoax content, legitimate website files, and configuration data.
  - Database Contents: Relevant databases containing hoax-content records or user interaction data.
- Acquisition Method Focus: Live acquisition using forensic tools (e.g., Oxygen Keyscout) to minimize service disruption and capture volatile data from a running system.

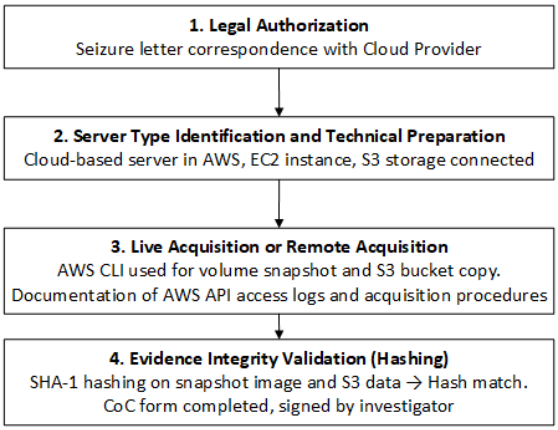


Fig. 6. Case 2 framework implementation.

Investigators received a report that a website hosted on a Virtual Private Server (VPS) was used to spread hoax news related to political information disturbing the public. Without stopping the server service, investigators needed to seize the web server and acquire related evidence, such as activity logs, site files, and hoax content posts. Simulation steps based on the proposed framework for Case 2 are shown in Fig. 6:

- Legal Authorization: Investigators begin by obtaining a court warrant to seize digital evidence from the VPS hosted web server. The warrant includes permission to retrieve site content files, user access logs, and databases that store hoax content and information related to user activity involved.
- Server Identification and Technical Preparation: Investigators identified that the web server was VPS and used Apache technology with PHP to run the site. Since the server was Virtual VPS, investigators prepared forensic tools such as Oxygen Keyscout and Magnet Axiom to perform live acquisition without shutting down the server. Investigators conduct an initial inspection of the server infrastructure, examine

the operating system (e.g., Linux) and prepare physical access to the server to collect evidence in real-time.

- Data Acquisition: Investigators perform live acquisition without stopping the web server service. Oxygen Keyscout is used to take snapshots of system files, including site content files, user activity logs, and databases that store hoax content. Live acquisition was performed using Oxygen Keyscout to capture system files associated with the website, including all files containing hoax news.
- Validation of Evidence Integrity: After the data is captured, investigators hash all captured files, logs, and databases to ensure the integrity of the evidence. After all files and access logs are retrieved, the hash values are calculated using Oxygen Keyscout and recorded in Chain of Custody to ensure that the collected data is not changed or manipulated.

The VPS simulation also completed live acquisition without downtime. Site files and logs were hash-validated and fully documented through chain-of-custody records, enabling accountable forensic analysis.

C. Metric

To assess the performance of the proposed framework, several metrics are defined based on technical, operational, and legal considerations. These metrics guide the evaluation process by providing quantifiable indicators of success throughout the simulated forensic procedures.

1) Aspects of success: This section outlines measurable aspects used to evaluate the effectiveness of the proposed framework in web server seizure scenarios. The framework is tested through two simulation scenarios: 1) seizure of a cloud web server used for online gambling and 2) seizure of a VPS web server used to spread hoax content. Each framework stage has specific success indicators, as follows:

- Volatile Data Preservation: Most of the $>$80% volatile data successfully preserved (RAM, active logs).
- Integrity Validation: Cryptographic hashing (SHA-1) performed pre- and post-acquisition to verify unchanged evidence values.
- Reporting and Secure Storage: The final report includes logs, hashes, analysis, and summary of findings. Evidence is stored encrypted and warrants are documented before acquisition.
- Downtime/Service Disruption: Minimal service disruption through live acquisition techniques.
- Cloud Environment Compatibility: Supports remote acquisition via official cloud provider APIs and services.
- Legal Compliance: Strict adherence to legal authorization and digital evidence handling standards.
- Ease of Analysis: Richer data set enables more thorough forensic analysis.

2) Quantitative evaluation results: To reduce subjectivity, the framework was also measured using quantitative indicators from both simulation cases (See Table II).

TABLE II. QUANTITATIVE RESULTS ACROSS SIMULATION SCENARIOS

Metric	Case 1 (Cloud)	Case 2 (VPS)
Hash verification success rate	100%	100%
Volatile artifact preservation rate	$\geq$80%	$\geq$80%
Acquisition completion time	$\approx$30 minutes	$\approx$30 minutes
Service downtime during acquisition	0 minutes	0 minutes
CoC completeness	$\geq$85%	$\geq$85%
Legal checklist compliance	$\geq$90%	$\geq$90%

Both scenarios satisfy the minimum feasibility target: average score $\geq$ 4.0 for all indicators and no critical legal or integrity indicator below 3.0.

3) Questionnaire: This section explains the questionnaire and interview guide used to evaluate the framework from practical and theoretical perspectives. The questionnaire includes questions related to:

- Preservation of digital evidence aspect
  - What is This framework provides systematic and standardized steps in seizure and acquisition of web servers?
  - Have the steps in this framework considered the preservation of volatile and non-volatile data with methods that are in line with forensic principles?
  - Does this framework provide an efficient cloud remote-acquisition method?
- Robust integrity and validation aspect
  - Does it provide evidence integrity validation mechanisms such as hashing to ensure the authenticity of the data obtained?
  - How well does it handle Chain of Custody to ensure evidence remains valid in court?
  - Is the framework flexible to be deployed on both VPS and cloud-hosted servers?
- Forensic acquisition admissibility in court aspect
  - To what extent does it meet legal requirements regarding the validity of obtaining evidence from seized web servers?
  - Does it meet the legal requirements regarding the seizure and retrieval of digital evidence, including the use of court warrants?

- What do you think are the weaknesses of this framework?

To assess the framework's feasibility, a structured questionnaire was developed and distributed to evaluators. This subsection describes the scoring criteria and rubric used to determine whether the framework meets minimum standards and legal admissibility. The framework is considered feasible if the average score for each indicator is ≥ 4.0 out of 5, and no critical aspects (legal and integrity) are scored ≤ 3, based on the Likert-scale rubric [49] below.

- (Very Poor): Evidence largely lost or corrupted; Weak validation and documentation; High inadmissibility.
- (Poor): Mostly data corrupted; Less validation, inconsistent documentation; May complicate admissibility.
- (Fair): Sufficient, but requires improvement and Sufficient to meet standards, but needs customization.
- (Good): Properly preserved using legitimate methods; Strong validation, detailed documentation.
- (Very Good): Comprehensive preserved, including live acquisition; Highly robust, transparent, and irrefutable documentation; Aligned with highest legal standards, beyond reasonable doubt.

4) Profile of SMEs: This section provides an overview of the Subject Matter Experts (SMEs) involved in the framework evaluation. Their professional backgrounds ensure a valid and informed assessment of the framework's applicability in real-world digital forensic contexts. The framework validation used an expert-judgment approach involving experts with direct experience in digital forensics [50]. Validation was conducted using both qualitative and quantitative assessment in two scenarios (cloud and VPS), involving 3 SMEs. Selection criteria included: 1) minimum 5 years of professional experience, 2) direct involvement in digital evidence handling, and 3) familiarity with legal admissibility requirements in Indonesia.

Feedback was collected through a structured questionnaire (Likert scale 1–5), guided interview, and review session on simulation outputs. The synthesis process applied indicator-level averaging, followed by cross-SME comparison to identify convergent findings and unresolved concerns.

A framework is categorized as feasible when the average score for each indicator is ≥ 4.0 and no critical indicator (legal and integrity aspects) is ≤ 3.0.

The SMEs were:

- Digital Forensics Practitioner
- Chairman of the Indonesian Digital Forensics Association
- Digital Forensic Expert at Indonesian Corruption Eradication Commission

Indonesian Corruption Eradication Commission (known as Komisi Pemberantasan Korupsi (KPK)) is a state institution established with the aim of improving the effectiveness and efficiency of efforts to eradicate corruption [51].

V. Results and Discussion

The development and evaluation results show the importance of aligning technical procedures with legal and operational standards in cloud and virtualized environments. The framework addresses gaps in existing models, especially in volatile-data preservation, integrity assurance, and minimal service disruption during live acquisition.

Integration of literature review and SME input strengthens the framework's legal validity and operational applicability. SME feedback confirms compliance with Indonesian legal standards and feasibility across VPS and cloud environments.

Simulation-based experimentation provides a replicable way to assess framework performance. Evaluation metrics were positive, with no critical SME failures, indicating that the framework can guide legally sound and technically reliable web server seizures.

A. SME Validation Results

This section discusses the evaluation results obtained from Subject Matter Experts (SMEs) who assessed the framework based on their professional judgment, practical experience, and legal-technical considerations. Their feedback serves as a crucial measure to validate the feasibility, effectiveness, and reliability of the proposed framework in real-world forensic investigations.

1) SME 1: Digital forensics practitioner: Validation focused on three aspects derived from the research objectives:

- Preservation of digital evidence aspect (Score: 4) The expert acknowledged that the framework already includes systematic and standardized procedures that align with forensic best practices. It successfully addresses the need to capture both volatile and non-volatile data, particularly by supporting live acquisition and remote access in cloud environments.
- Robust integrity and validation aspect (Score: 4) The SME confirmed that the framework incorporates proper integrity validation mechanisms, notably hashing (e.g., SHA-256), and emphasizes the importance of documenting each acquisition step. Chain of Custody was positively reviewed as well-structured and able to meet legal requirements.
- Forensic acquisition admissibility in court aspect (Score: 4) The legal compliance elements such as prior authorization, use of court warrants, and procedural transparency were considered robust. The SME stated that the framework is legally defensible and suitable for submission as digital evidence in court.

2) SME 2: Chairman of the Indonesian digital forensics association:

- Preservation of digital evidence aspect (Score: 4.67) The expert stated that the proposed framework already has systematic steps for collecting electronic evidence. However, as input, the expert provided notes regarding detailed elaboration on data preservation and acquisition, as well as technical steps related to remote acquisition.

- Robust integrity and validation aspect (Score: 4.67) The proposed framework is considered excellent in terms of integrity and strong evidence validation. This framework provides a hashing mechanism to ensure the authenticity and integrity of the data obtained, and has incorporated clear Chain of Custody procedures to ensure that evidence remains valid and accountable in court. Although it is flexible for use on VPS and cloud servers, further elaboration is needed regarding the different types for technical acquisition details.
- Forensic acquisition admissibility in court aspect (Score: 4.67) The framework is considered to meet both legal and technical requirements for evidence from web server seizures to be admissible in court. Legally, it emphasizes obtaining court authorization for seizure, aligning it with existing procedures for physical evidence under KUHAP.

3) SME 3: Digital forensic expert at Indonesian corruption eradication commission:

- Preservation of digital evidence aspect (Score: 4.33) The expert stated that the proposed framework is considered highly systematic and can serve as a reference for web server seizure, although there a note to strengthen documentation for live RAM and volatile logs. Additionally, the framework provides efficient methods for remote acquisition on cloud servers.
- Robust integrity and validation aspect (Score: 4.67) This framework provides robust mechanisms for validating evidence integrity, such as hashing, which is included in the standard procedure for digital evidence collection. The handling of the Chain of Custody is rated as very good because documentation is prepared from the beginning of the acquisition process to ensure the validity of evidence in court.
- Forensic acquisition admissibility in court aspect (Score: 4.67) This framework is considered compliant with legal requirements for the admissibility of evidence from web server seizures in court, provided that procedures are carried out according to seizure and hashing regulations.

B. Framework Feasibility

The variation in SME scores can be explained by the following factors:

- Diverse Backgrounds and Expertise: SMEs come from different fields such as law enforcement, legal practice, and cybersecurity, leading to varied priorities.
- Interpretation of Framework Components: The subjective nature of the scoring process means that SMEs interpret the framework's components differently, resulting in discrepancies based on their individual expectations and focus areas.
- Variability in Case Scenarios: The applicability of the framework can differ based on the specific case scenarios encountered by SMEs, influencing their scores depending on their experiences with cloud-based versus traditional VPS environments.
- Feedback and Iterative Improvement: Score differences provide constructive input for framework refinement.

Overall, differing SME scores reflect complementary perspectives rather than contradiction, and they help identify priorities for improvement.

The framework provides a solid foundation for conducting digital forensic investigations in Indonesia. All critical aspects related to legal compliance and data integrity received scores above the minimum requirement, with no aspect falling below the threshold of 3.0 [49]. A score below 3.0 in these areas would indicate significant weaknesses that could jeopardize the admissibility of evidence in court.

The proposed framework meets the threshold criteria, with an average score of **4.41**/5.0, indicating strong SME support. These findings suggest that the framework can be implemented while continuing iterative refinement [52].

C. Chain of Custody in Cloud-Native Environments: Interim Procedure and Design Direction

Although the framework demonstrates strong legal and technical feasibility, cloud-native implementation requires additional optimization in chain-of-custody automation. As an interim procedure, this study recommends:

- Generate immutable timestamped acquisition logs for each command/API action.
- Apply digital signatures to evidence manifests and hash reports.
- Store forensic artifacts and custody logs in write-once object storage (WORM/object-lock policy).
- Record actor identity and role for every evidence access event.
- Correlate cloud audit records (e.g., CloudTrail) with local acquisition notes before final reporting.

For future design, a cloud-native chain-of-custody module can be developed by integrating automated event capture, cryptographically linked audit trails, and policy-based verification to support end-to-end courtroom defensibility.

VI. Conclusion

This study addresses the need for a standardized web server seizure framework in Indonesia by integrating legal requirements with practical forensic procedures for VPS and cloud environments. The framework was developed through literature review, expert input, simulation, and refinement.

Results show that the framework supports legally authorized seizure, scenario-based technical preparation, live or remote acquisition, integrity validation through hashing, and complete chain-of-custody documentation. Across the simulations, evidence integrity and legal compliance were maintained without service downtime.

SME evaluation indicates strong feasibility (average score **4.41**/5.0), especially in legal compliance, integrity assurance, and acquisition effectiveness. The framework is therefore suitable as a practical reference, with future work focused on cloud-native chain-of-custody automation.

ACKNOWLEDGEMENTS

Thanks to the Subject Matter Experts who contributed their valuable insights and expertise throughout this research. Their constructive feedback and professional evaluations played a pivotal role in validating the proposed framework and ensuring its practical applicability within the field of digital forensics.

REFERENCES

[1] B. S. dan Sandi Negara, "Lanskap keamanan siber indonesia 2024," https://www.bssn.go.id/monitoring-keamanan-siber/, 2024, accessed: 2025-02-07.

[2] F. Thealma and Y. Ruldeviyani, "Digital forensic ethical data handling in indonesia," *The Indonesian Journal of Computer Science*, vol. 14, 2 2025. [Online]. Available: http://www.ijcs.net/ijcs/index.php/ijcs/article/view/4588

[3] R. Stoykova, "The right to a fair trial as a conceptual framework for digital evidence rules in criminal investigations," *Computer Law & Security Review*, vol. 49, p. 105801, 7 2023.

[4] W. Hartono, D. Muhardi, A. Akhiruddin, D. Valentianna, B. Purba, P. Asa, and Y. Dm, "Challenges of criminal investigation cyber crime," *Awang Long Law Review*, vol. 7, pp. 11–19, 11 2024. [Online]. Available: https://ejournal.stih-awanglong.ac.id/index.php/awl/article/view/1351

[5] G. Horsman and N. Sunde, "Unboxing the digital forensic investigation process," *Science & Justice*, vol. 62, pp. 171–180, 3 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1355030622000144

[6] F. Ramadhanny, "Aneka situs pemerintah dijaili jadi peringatan darurat garuda biru," *detik.com*, 2024. [Online]. Available: https://inet.detik.com/security/d-7505426/aneka-situs-pemerintah-dijaili-jadi-peringatan-darurat-garuda-biru

[7] Y. C. Tok, D. Y. Zheng, and S. Chattopadhyay, "A smart city infrastructure ontology for threats, cybercrime, and digital forensic investigation," *Forensic Science International: Digital Investigation*, vol. 52, p. 301883, 3 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281725000228

[8] A. Singh, "Operation power off," 2022. [Online]. Available: https://www.cyberpeace.org/resources/blogs/operation-power-off

[9] S. Kang, U. Hur, G. Kim, and J. Kim, "Forensic analysis and data decryption of tencent meeting in windows environment," *Forensic Science International: Digital Investigation*, vol. 51, p. 301818, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281724001422

[10] T. Göbel, F. Breitinger, and H. Baier, "Optimising data set creation in the cybersecurity landscape with a special focus on digital forensics: Principles, characteristics, and use cases," *Forensic Science International: Digital Investigation*, vol. 52, p. 301882, 3 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281725000216

[11] M. D. Kohn, M. M. Eloff, and J. H. Eloff, "Integrated digital forensic process model," *Computers and Security*, vol. 38, pp. 103–115, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.cose.2013.05.001

[12] F. Casino, T. K. Dasaklis, G. P. Spathoulas, M. Anagnostopoulos, A. Ghosal, I. Borocz, A. Solanas, M. Conti, and C. Patsakis, "Research trends, challenges, and emerging topics in digital forensics: A review of reviews," *IEEE Access*, vol. 10, pp. 25 464–25 493, 2022.

[13] T. J. Silva, E. OliveiraJr, M. E. Pereira, and A. F. Zorzo, "A review study of digital forensics in iot: Process models, phases, architectures, and ontologies," *Forensic Science International: Digital Investigation*, vol. 53, p. 301912, 6 2025. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2666281725000514

[14] S. M. Pedapudi and N. Vadlamani, "Data acquisition based seizure record framework for digital forensics investigations," *Proceedings of the 5th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2021*, pp. 1766–1768, 2021.

[15] Oxford, "seizure noun - definition, pictures, pronunciation and usage notes — oxford advanced learner's dictionary at oxfordlearnersdictionaries.com." [Online]. Available: https://www.oxfordlearnersdictionaries.com/definition/english/seizure

[16] T. R. Sree and S. M. S. Bhanu, "Data collection techniques for forensic investigation in cloud," *Digital Forensic Science*, 2020.

[17] S. Ruiz-Villafranca, J. M. C. Gómez, and J. Roldán-Gómez, "A forensic tool for the identification, acquisition and analysis of sources of evidence in iot investigations," *Internet of Things*, vol. 27, p. 101308, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S254266052400249X

[18] C. Liu, A. Singhal, and D. Wijesekera, "Identifying evidence for cloud forensic analysis," in *IFIP Advances in Information and Communication Technology*, vol. 511. Springer New York LLC, 2017, pp. 111–130.

[19] H. Spichiger and F. Adelstein, "Preserving meaning of evidence from evolving systems," *Forensic Science International: Digital Investigation*, vol. 52, p. 301867, 2025, dFRWS EU 2025 - Selected Papers from the 12th Annual Digital Forensics Research Conference Europe. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266628172500006X

[20] J. P. A. Yaacoub, H. N. Noura, O. Salman, and A. Chehab, "Advanced digital forensics and anti-digital forensics for iot systems: Techniques, limitations and recommendations," *Internet of Things*, vol. 19, p. 100544, 8 2022. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2542660522000464

[21] V. Prakash, A. Williams, L. Garg, P. Barik, and R. K. Dhanaraj, "Cloud-based framework for performing digital forensic investigations," *International Journal of Wireless Information Networks*, vol. 29, pp. 419–441, 12 2022. [Online]. Available: https://link.springer.com/article/10.1007/s10776-022-00560-z

[22] A. M. Komarudin, N. Widiyasono, A. P. Aldya, and R. Rizal, "Data integrity testing of digital evidence data capture results on private cloud computing services," *INNOVATICS: Innovation in Research of Informatics*, vol. 5, 11 2023. [Online]. Available: https://jurnal.unsil.ac.id/index.php/innovatics/article/view/8420

[23] A. Almuqren, H. Alsuwaelim, M. M. H. Rahman, and A. A. Ibrahim, "A systematic literature review on digital forensic investigation on android devices," *Procedia Computer Science*, vol. 235, pp. 1332–1352, 1 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924008020

[24] B. Sharma, J. Ghawaly, K. McCleary, A. M. Webb, and I. Baggili, "Forensicllm: A local large language model for digital forensics," *Forensic Science International: Digital Investigation*, vol. 52, p. 301872, 3 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281725000113

[25] J. Yang, J. Kim, J. Bang, S. Lee, and J. Park, "Catch: Cloud data acquisition through comprehensive and hybrid approaches," *Forensic Science International: Digital Investigation*, vol. 43, p. 301442, 9 2022.

[26] R. L. Lau, "Web server part 1: Apache/nginx basics," *Practical Internet Server Configuration*, pp. 183–225, 2021. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4842-6960-2_9

[27] R. U. Rahman and D. S. Tomar, "A new web forensic framework for bot crime investigation," *Forensic Science International: Digital Investigation*, vol. 33, p. 300943, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281720300718

[28] ISO, "Iso 27037:2012 - guidelines for identification, collection, acquisition and preservation of digital evidence," 2012. [Online]. Available: https://www.iso.org/standard/44381.html

[29] D. Day, "Chapter 7 - seizing, imaging, and analyzing digital evidence: step-by-step guidelines," in *Cyber Crime and Cyber Terrorism Investigator's Handbook*, B. Akhgar, A. Staniforth, and F. Bosco, Eds. Syngress, 2014, pp. 71–89. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128007433000074

[30] A. Wickramasekara, F. Breitinger, and M. Scanlon, "Exploring the potential of large language models for improving digital forensic investigation efficiency," *Forensic Science International: Digital Investigation*, vol. 52, p. 301859, 3 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281724001860

[31] C. Karagiannis and K. Vergidis, "Digital evidence and cloud forensics: Contemporary legal challenges and the power of disposal," *Information 2021, Vol. 12, Page 181*, vol. 12, p. 181, 4 2021. [Online]. Available: https://www.mdpi.com/2078-2489/12/5/181/htmhttps://www.mdpi.com/2078-2489/12/5/181

[32] S. V. Patel and I. S. Lurie, "The use of portable separation devices for forensic analysis: A review of recent literature," *Forensic Chemistry*, vol. 26, p. 100365, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468170921000618

[33] K. U. Maheswari and G. Shobana, "The state of the art tools and techniques for remote digital forensic investigations," *2021 3rd International Conference on Signal Processing and Communication, ICPSC 2021*, pp. 464–468, 5 2021.

[34] K. Kent, S. Chevalier, T. Grance, and H. Dang, "Special publication 800-86 guide to integrating forensic techniques into incident response recommendations of the national institute of standards and technology," National Institute of Standards and Technology, Tech. Rep., 2006.

[35] K.-S. Lim, A. Savoldi, C. Lee, and S. Lee, "On-the-spot digital investigation by means of ldfs: Live data forensic system," *Mathematical and Computer Modelling*, vol. 55, no. 1, pp. 223–240, 2012, advanced Theory and Practice for Cryptography and Future Security. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895717711002895

[36] A. W. Malik, D. S. Bhatti, T. J. Park, H. U. Ishtiaq, J. C. Ryou, and K. I. Kim, "Cloud digital forensics: Beyond tools, techniques, and challenges," *Sensors 2024, Vol. 24, Page 433*, vol. 24, p. 433, 1 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/2/433/htmhttps://www.mdpi.com/1424-8220/24/2/433

[37] J.-P. Sandvik, K. Franke, H. Abie, and A. Årnes, "Quantifying data volatility for iot forensics with examples from contiki os," *Forensic Science International: Digital Investigation*, vol. 40, p. 301343, 2022, selected Papers of the Ninth Annual DFRWS Europe Conference. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281722000129

[38] W. Yun, J. Kang, S. Lee, and J. Park, "Digital forensic approaches to intel and amd firmware raid systems," *Forensic Science International: Digital Investigation*, vol. 54, p. 301971, 9 2025. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2666281725001106

[39] N. Kuntze, C. Rudolph, H. Schilling, A. Alva, B. Brisbois, and B. Endicott-Popovsky, "Seizure of digital data and 'selective suppression' of digital evidence," *Int. Workshop Syst. Approaches Digit. Forensics Eng., SADFE*, 9 2014.

[40] J. G. J. Nortje and D. C. Myburgh, "Impediments during the compilation of a search and seizure warrant for digital information by forensic investigators in south africa," *Journal of Financial Crime*, vol. 31, pp. 476–488, 4 2024.

[41] K. S. Singh, A. Irfan, and N. Dayal, "Cyber forensics and comparative analysis of digital forensic investigation frameworks," *2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019*, pp. 584–590, 2019.

[42] C. Hargreaves, F. Breitinger, L. Dowthwaite, H. Webb, and M. Scanlon, "Dfpulse: The 2024 digital forensic practitioner survey," *Forensic Science International: Digital Investigation*, vol. 51, p. 301844, 12 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281724001719

[43] B. Manan, "Prosedur penyitaan." [Online]. Available: https://pn-metro.go.id/index.php/tentang-pengadilan/2019-09-30-06-51-00/pidana/prosedur-penyitaan

[44] D. B. Andersen, N. Sunde, and K. Porter, "Tool induced biases? misleading data presentation as a biasing source in digital forensic analysis," *Forensic Science International: Digital Investigation*, vol. 52, p. 301881, 3 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281725000204

[45] T. J. Silva, A. H. Mazur, E. OliveiraJr, A. F. Zorzo, and M. P. Barcellos, "An ontology for promoting controlled experimentation in digital forensics," *Forensic Science International: Digital Investigation*, vol. 52, p. 301845, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281724001720

[46] D. Rani, N. S. Gill, and P. Gulia, "A forensic framework to improve digital image evidence administration in iiot," *Journal of Industrial Information Integration*, vol. 38, p. 100568, 3 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2452414X24000128

[47] B. Martini and K. K. R. Choo, "An integrated conceptual digital forensic framework for cloud computing," *Digital Investigation*, vol. 9, pp. 71–80, 11 2012. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S174228761200059X\#sec3.3

[48] R. Stoykova and K. Franke, "Reliability validation enabling framework (rvef) for digital forensics in criminal investigations," *Forensic Science International: Digital Investigation*, vol. 45, p. 301554, 6 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266628172300063X

[49] J. Robinson, "Likert scale," *Encyclopedia of Quality of Life and Well-Being Research*, pp. 3917–3918, 2023. [Online]. Available: https://link.springer.com/rwe/10.1007/978-3-031-17299-1_1654

[50] N. Martynenko, "Digitalisation of forensic expert activity in ukraine: Organisational and legal framework," *Forensic Science International: Synergy*, vol. 10, p. 100578, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589871X25000075

[51] Undang-Undang, "Undang-undang (uu) nomor 19 tahun 2019 tentang perubahan kedua atas undang-undang nomor 30 tahun 2002 tentang komisi pemberantasan tindak pidana korupsi," Meteri Hukum dan Hak Asasi Manusia, Tech. Rep., 2019. [Online]. Available: www.peraturan.go.id

[52] M. Ng, J. James, and R. Bull, ""what you say in the lab, stays in the lab": A reflexive thematic analysis of current challenges and future directions of digital forensic investigations in the uk," *Forensic Science International: Digital Investigation*, vol. 51, p. 301839, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281724001665

Adaptive Phishing Website Detection Using Incremental Machine Learning: A Dynamic Approach to Cybersecurity Threats

Ajla Kulaglic, Mutaz A. B. Al-Tarawneh
College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait

***Abstract*—The rapid expansion of internet services and cloud-based platforms has increased cybersecurity threats, particularly phishing attacks that deceive users into disclosing sensitive information. Traditional phishing detection methods, including blacklists and batch-learning models, often struggle to adapt to the continuously evolving nature of these attacks. In order to address this challenge, this study proposes an adaptive phishing detection framework based on incremental machine learning techniques that enable real-time learning and dynamic adjustment to new attack patterns. A comprehensive evaluation of multiple incremental algorithms was performed using the River-ML framework and a publicly available phishing website dataset. The models were assessed based on accuracy, precision, recall, F1 score, Cohen's kappa, and memory efficiency. Evaluation results demonstrate that models such as Aggregated Mondrian Forest, Extremely Fast Decision Trees, and Logistic Regression achieved strong classification performance, with the best accuracy reaching 90.15%, precision up to 91.05%, recall up to 89.42%, F1 score up to 88.75%, and Cohen's kappa up to 79.99%, while lightweight models like ALMA maintained extreme memory efficiency, requiring as little as 1.81 KB. In general, the proposed incremental learning framework significantly improves the effectiveness of phishing detection and computational efficiency, providing a scalable and adaptive defense mechanism against evolving cyber threats.**

***Keywords*—Phishing detection; incremental learning; online machine learning; cybersecurity threats; real-time classification**

I. Introduction

Phishing is a deceptive cyber attack that exploits both social engineering and technical manipulation to unlawfully obtain user credentials and financial information. Attackers craft fraudulent emails, messages, or website links that closely resemble legitimate communications from reputable organizations, luring users to access counterfeit websites where they unknowingly disclose sensitive details such as usernames and passwords. Beyond deceptive messaging, hackers further amplify these attacks by deploying malicious software designed to extract login credentials and intercept user data from compromised systems. Various techniques are used, including email spoofing, manipulated URLs, instant messaging, forum postings, phone calls, and SMS phishing (smishing) to trick users into providing personal information. The authenticity of phishing content often mirrors that of genuine platforms, making it difficult for users to distinguish between legitimate and fraudulent interactions. The primary objective of phishing attacks is to obtain personal data for financial gain or identity theft, causing significant economic damage on a global scale [1].

Cybercriminals acquire confidential data by replicating legitimate websites and emails, often impersonating financial institutions or organizations that handle sensitive financial transactions [2], [3], [4]. These fraudulent emails are meticulously designed using authentic company logos, slogans, and branding elements to appear genuine. By exploiting HTML structures, attackers replicate entire websites or images, making them visually indistinguishable from their legitimate counterparts [5]. This deceptive practice has been facilitated by the widespread use of the Internet as a primary communication medium, allowing threat actors to exploit well-known brands, trademarks, and corporate identities that users trust for verification. To maximize their reach, phishers distribute large volumes of deceptive emails, commonly referred to as "spooled" emails, targeting unsuspecting recipients. Once these emails are opened, victims are typically redirected from an authentic-looking platform to a fraudulent website, where they unknowingly submit personal credentials [6], [7], [8]. The exploitation of user information through phishing has become a critical cybersecurity challenge in modern society [9], [10], [11]. Extensive research has been conducted to analyze various phishing indicators, including website URLs, webpage content, source code structures, and visual similarities, in an effort to mitigate these threats [12], [13], [14].

However, a significant gap remains in the availability of adaptive and efficient anti-phishing tools that can proactively detect and counteract evolving phishing tactics. Machine Learning (ML) techniques have been proven to be highly effective in identifying malicious URLs in real-time, offering a more dynamic alternative to conventional detection methods [15], [16], [17], [18], [19], [20], [21]. Traditional blacklist-based URL detection relies on manually curated databases of known phishing links, which require frequent updates and struggle to keep pace with the continuous emergence of new phishing domains. Attackers further evade detection through automated techniques such as Domain Generation Algorithms (DGA), rendering blacklist-based defenses inadequate [22]. The complexity of maintaining an exhaustive blacklist highlights the need for more intelligent and adaptive detection mechanisms. Researchers have increasingly advocated for ML-driven approaches, which leverage pattern recognition and predictive modeling to classify URLs as malicious or benign. By framing phishing detection as a binary classification problem, ML models can efficiently map high-dimensional feature spaces to corresponding threat labels, offering superior adaptability and generalization compared to conventional blacklist-based methods [23]. Despite growing interest in ap-

plying machine learning techniques for phishing detection, most existing approaches rely on batch or offline learning, which may be insufficient to handle evolving and dynamic attack strategies. Batch learning models present several limitations, particularly in the context of phishing detection. First, the training process is computationally intensive, requiring substantial processing power and memory resources. Second, the performance of the model is highly dependent on the size and quality of the training data, making generalization to new threats a challenge. Third, batch learning operates under the assumption that training data remain static over time, preventing the model from adapting to new samples or changes in data distribution (concept drift), necessitating frequent retraining and model redevelopment. Given the continuous nature of security threats and the dynamic landscape of web activity, incremental machine learning algorithms offer a more effective alternative to batch-based approaches. These algorithms are designed to process unbounded data streams, allowing continuous learning and real-time model updates. Additionally, they are well-suited for real-time phishing detection, addressing challenges that traditional batch models struggle with, such as handling rapidly evolving threats. In addition, incremental learning excels at detecting concept drift, enabling models to dynamically adjust to changing attack patterns and browsing behaviors. This adaptability makes incremental machine learning a promising solution for phishing detection in environments where threat patterns change unpredictably over time [24], [25], [26]. Several incremental machine learning algorithms have been proposed in the literature [27], [28], [29], [26], [30], [31], [32], [33], yet their applicability to phishing detection and the trade-offs between performance metrics and memory efficiency remain underexplored. This study aims to systematically evaluate various incremental learning algorithms to determine their effectiveness in detecting phishing attempts. In addition, an in-depth analysis of key performance trade-offs, including accuracy, precision, recall, kappa statistic, and memory consumption, is conducted to identify the most suitable models for real-time phishing detection.

The remainder of this study is structured as follows: Section II reviews related research efforts in phishing detection and incremental learning. Section III outlines the research methodology and details the experimental setup and evaluation framework. Section IV presents the results of the online machine learning evaluation and examines the effectiveness of online model selection using multi-armed bandits. Finally, Section V summarizes the findings and concludes the study.

II. Literature Review

Phishing is a cyberattack in which malicious actors impersonate reputable websites or businesses to deceive users into revealing sensitive personal information, such as passwords, usernames, and bank account details. These attacks exploit human trust and employ deceptive strategies, including social engineering, email spoofing, and URL manipulation, to mislead unsuspecting users. Phishing incidents can have severe consequences, including financial loss, identity theft, and compromised cybersecurity. Various detection techniques have been proposed in the literature to counteract these threats, including List-Based, Visual Similarity, Heuristic, Machine Learning, and Deep Learning approaches [34].

List-based detection techniques are widely employed in web browsers such as Google Chrome, Microsoft Edge, and Firefox to identify phishing websites [35]. These approaches rely on whitelisting and blacklisting methods. A whitelist contains a list of legitimate URLs that browsers are allowed to access, ensuring that only approved websites can be loaded. In contrast, a blacklist maintains a database of known phishing or fraudulent URLs, preventing browsers from loading potentially malicious web pages. Although blacklisting can effectively block previously identified phishing websites, it requires frequent updates to remain effective. Even minor modifications to a phishing URL can bypass blacklist-based defenses, making this method less reliable against newly emerging threats.

Visual similarity-based detection methods aim to identify phishing websites by analyzing their similarities to legitimate ones. This approach compares various visual elements, including the logo of the website, text layout, CSS, source code, and screenshots, to determine potential fraud. Since phishing websites often mimic the appearance of legitimate platforms, this technique is useful for detecting impersonation attempts. However, it has limitations in identifying zero-hour phishing attacks, as it relies on previously recorded or stored websites for comparison. Without prior knowledge of new phishing attempts, visual similarity-based detection may fail to recognize novel threats [36]. Heuristic-based detection methods analyze various characteristics of a website to distinguish between phishing and legitimate web pages. This approach examines several features, including URL structures, domain registration details, SSL certificate status, website traffic, and DNS records. Using predefined rules and classification algorithms, heuristic detection can identify phishing attempts even when the fraudulent website is not included in a blacklist. The effectiveness of this method depends on the quality of the selected set of features, training samples, and classification techniques. One significant advantage of heuristic-based detection is its ability to identify zero-hour phishing attacks, making it a more dynamic approach compared to static detection mechanisms.

Machine learning techniques have gained widespread attention for phishing detection due to their ability to learn patterns from large datasets and identify emerging threats.With the increasing availability of large-scale datasets, machine learning has proven to be an efficient tool in identifying phishing websites, especially in high-speed and high-volume data environments [37]. Machine learning models are trained using features extracted from phishing URLs, website structures, and JavaScript behavior. These extracted attributes help classifiers, such as Support Vector Machines (SVM), Random Forest, and Neural Networks, differentiate between phishing and benign websites. By analyzing patterns in big data environments, machine learning-based detection methods provide improved accuracy and adaptability over traditional list-based approaches. However, the effectiveness of machine learning models depends on the quality of the data and the selection of features [38]. Deep learning techniques further enhance phishing detection by leveraging advanced neural networks to automatically extract complex features from web pages. Recent studies suggest that deep learning models can outperform traditional machine learning classifiers in detecting phishing websites. Popular deep learning architectures for phishing detection include deep neural networks (DNN), convolutional neural networks (CNN), recurrent neural networks (RNN),

feed-forward deep neural networks, restricted Boltzmann machines, deep belief networks and deep autoencoders [39]. These models can analyze web content, user behavior, and URL characteristics to identify phishing attempts with higher accuracy. The ability of deep learning models to generalize patterns and adapt to new phishing techniques makes them a promising solution to combat evolving cybersecurity threats. Despite the effectiveness of existing detection methods, phishing attacks continue to evolve, requiring more adaptive and real-time detection mechanisms. The use of incremental learning in phishing detection remains an underexplored area, highlighting a critical gap in research. Addressing this limitation by incorporating incremental learning techniques could improve the robustness and efficiency of phishing detection systems in dynamic environments. The remainder of this section provides a more detailed discussion of machine learning and deep learning-based techniques, as they play a crucial role in phishing detection, particularly in the approach proposed in this study.

The study in [40] extensively examined the application of machine learning techniques to URL-based phishing detection. The authors evaluated multiple algorithms, including Naïve Bayes, Random Forest, K-Nearest Neighbor, AdaBoost, K-star, Support Vector Machines, and Decision Trees, using a self-constructed phishing dataset. Their findings indicated that the Random Forest algorithm, which is based solely on NLP-based features, exhibited the highest performance, achieving an accuracy rate of 97.98%. These results underscore the effectiveness of feature engineering to improve phishing detection. Similarly, the research conducted in [41] assessed the classification capabilities of Support Vector Machines, Decision Trees, and Logistic Regression using the PhishTank and DMOZ phishing datasets. Their findings revealed that among the models tested, the Support Vector Machines achieved the highest accuracy of 89.3%, demonstrating their robustness in distinguishing phishing from legitimate websites.

Further expanding on machine learning models, the study in [42] explored the performance of Random Forest, Support Vector Machines, Naïve Bayes, C4.5, JRip and PART classifiers in several phishing datasets, including PhishTank, OpenPhish, Alexa and Common Crawl. The results demonstrated that the Random Forest classifier outperformed other models, achieving an accuracy of 96.17%. This reinforces the previously established effectiveness of Random Forest in phishing detection. Based on these findings, [43] investigated the classification of legitimate and illegitimate websites using machine learning models such as Random Forest, Decision Trees, K-Nearest Neighbor (KNN), Iterative Dichotomiser-3 (ID3), and Naïve Bayes, trained on the UCI phishing dataset. Their study highlighted the importance of feature selection, as Genetic Algorithms (GAs) were shown to improve detection accuracy when used for feature optimization. More specifically, the combination of Iterative Dichotomiser-3 (ID3) with Yet Another Generating Genetic Algorithm (YAGGA) led to a significant increase in detection accuracy, reaching up to 95%. In [44], a novel approach was introduced, in which the authors developed the Hybrid Feature-Based Phishing Classifier (PHFBC), which integrates the Decision Tree and Naïve Bayes models with a statistical measure called the phishing ratio. Furthermore, the study proposed a robust feature selection technique using the Recursive Feature Subset Selection Algorithm (RFSSA). The proposed methodology was tested on the PhishTank, Chinese e-Business and DMOZ phishing datasets, achieving a True Positive Rate (TPR) of 0.984. This further emphasizes the importance of feature selection in improving detection performance. In another study, [45] focused on optimizing machine learning-based phishing detection through feature selection and hyperparameter tuning. The research explored multiple optimization strategies, including the Tree-structured Parzen Estimator (TPE) and Genetic Algorithms (GA) for hyperparameter optimization, as well as Moth Flame Optimization (MFO) and Particle Swarm Optimization (PSO) for feature selection. The experiments conducted on the PhishTank and Alexa datasets indicated that the optimal combination was Random Forest with PSO for feature selection and TPE for hyperparameter tuning, resulting in an impressive accuracy of 99.33%. These findings highlight the critical role of optimization techniques in improving the accuracy of phishing detection. In a related effort, [46] proposed a phishing detection system that uses URL analysis in eight different machine learning algorithms, including Logistic Regression, K-Nearest Neighbor, Decision Tree, Support Vector Machines, Naïve Bayes, XGBoost, Random Forest, and Artificial Neural Networks. Their comparative analysis on three different datasets—PhishTank, Alexa, and Common Crawl—showed that Random Forest applied to the PhishTank dataset achieved the highest accuracy of 94.59%. This reinforces the dominance of Random Forest in phishing detection applications. Another significant study was conducted by [47], who explored the use of domain name features for phishing detection by incorporating DNS data, blacklists, and lexical attributes. Their classifier, based on logistic regression, was trained using four datasets: PhishTank, Alexa, ICANN, and DNS-BH. However, while their approach demonstrated effectiveness in distinguishing between benign and malicious domains, overall accuracy remained relatively low at approximately 60%, indicating the limitations of purely lexical-based phishing detection. In contrast, [48] introduced a data augmentation method to improve machine learning-based phishing detection. Their study leveraged an Adversarial Autoencoder (AAE) to generate synthetic phishing samples, enhancing training datasets. The proposed model was tested on multiple classifiers, including support vector machines, decision trees, gradient booster, K-nearest neighbor, and random forest, using the UCI and Mendeley phishing datasets. The results demonstrated that models trained with AAE-generated data exhibited greater robustness and accuracy. Specifically, the Gradient Boosting classifier achieved the highest accuracy of 95.47%, validating the benefits of augmenting phishing datasets with adversarial examples. Expanding on previous research, the study in [49] examined the structural characteristics of URLs from phishing websites by extracting 12 different types of information. The authors trained four machine learning algorithms, namely Logistic Regression, Support Vector Machine, Naïve Bayes, and Decision Tree, using the PhishTank and DMOZ datasets. Their findings indicated that Logistic Regression outperformed the other classifiers, achieving an optimal accuracy of 95.12%. These results highlight the importance of URL-based feature extraction in enhancing the accuracy of phishing detection. Further advancing the field, the authors in [50] tackled the challenge of phishing detection within e-commerce platforms by proposing a framework for constructing reproducible and extensible datasets. Their methodology involved categorizing and selecting 87 widely

recognized phishing features, assessing their relevance and runtime performance, and leveraging them to build a robust dataset. Through conceptual replication and evaluation across the PhishTank, Alexa, and OpenPhish datasets, their results demonstrated that Random Forest was the most effective classifier, particularly when utilizing features from external services. Furthermore, their analysis showed that hybrid features contributed to the highest classification accuracy of 96.61%, while filter-based feature selection methods outperformed wrapper techniques, reaching an accuracy of 96.83%. Similarly, the work in [51] introduced a novel approach to detect real-time phishing attacks by evaluating deep hybrid features using the Light Gradient Boosted Machine model. Their method involved normalizing web page request features and applying Sparse Autoencoder and Principal Component Analysis for dimensionality reduction. The experimental evaluation, conducted on the ISCXURL-2016 dataset, demonstrated that their approach achieved a high classification accuracy of 99.6%, effectively minimizing false positives before processing web requests.

Extending the discussion on phishing datasets, the study in [52] highlighted the limitations of existing datasets, particularly their exclusion of login pages, which are critical for phishing detection. To address this gap, the authors introduced PILU-60K, a novel dataset that incorporates login URLs. Their experimental analysis revealed that while the support vector machines exhibited resilience against evolving phishing tactics, Random Forest achieved superior performance with PILU-60K, achieving an accuracy of 94.59%. These findings underscore the need to include login page data in phishing datasets to enhance detection accuracy. Furthermore, the research in [53] examined the security risks associated with browsing the Internet, particularly due to the low awareness of users of cyber threats such as phishing attacks. Their study introduced a supervised machine learning approach that utilizes Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine, and Multi-Layer Perceptron classifiers. By extracting novel features solely from URLs and testing them on the PhishTank dataset, the study assessed the effectiveness of their approach compared to Google Safe Browsing (GSB), the default security control in major web browsers. Their results demonstrated that the proposed method consistently outperformed GSB, even when tested against phishing URLs that remained active for more than a year after model training. In particular, their optimized Random Forest model achieved the highest accuracy level of 99.29%. Building on the effectiveness of ensemble learning, [54] introduced PDCFS, a novel phishing detection model that uses hybrid cumulative feature selection. Their model partitioned datasets into multiple subsets based on diverse feature selection techniques, such as Chi-Square and Principal Component Analysis, and employed classifiers like Support Vector Machine and Random Forest. By implementing a five-fold cross-validation strategy and aggregating the results through majority voting, their findings indicated that Random Forest, when applied to a reduced feature set of 32, achieved the highest accuracy of 98.36%. Meanwhile, the PDCFS model itself achieved a close accuracy of 98.24%, demonstrating its competitive performance against other hybrid models.

The study in [55] explored the importance of phishing attacks and the ongoing challenge of developing robust detection mechanisms, particularly against zero-day attacks. To address these issues, the authors proposed a novel approach that integrates convolutional operations to model character-level URL features with a deep convolutional autoencoder (CAE). Their extensive experiments, conducted on three real-world datasets—PhishTank, PhishStorm, and ISCX-URL-2016—spanning a total of 222,541 URLs, demonstrated that the proposed method outperformed the latest deep learning models. In particular, receiver operating characteristic (ROC) curve analysis and 10-fold cross-validation revealed a 3.98% improvement in sensitivity over existing state-of-the-art models. These results underscore the effectiveness of using both sustainability and intelligibility in phishing detection, reinforcing the importance of deep learning-driven feature representation. Similarly, the research presented in [56] examined the surge in cyber-attacks, particularly phishing, that intensified due to increased digital activity during the pandemic. The study implemented a phishing detection system based on Convolutional Neural Networks (CNNs), where n-gram features were extracted from URLs using the PhishTank dataset. Experimental findings indicated that unigram features, particularly those focused on 70 specified characters irrespective of case sensitivity, achieved the highest accuracy rate of 88.90%. Additionally, the system demonstrated efficient URL classification, processing each URL in approximately 0.008 seconds. This research highlights the potential of deep learning models in achieving fast and effective phishing detection. Expanding on hybrid approaches, [57] proposed an innovative spam detection model that integrates dataset acquisition, feature extraction, optimal feature selection, and detection phases. The study collected a benchmark dataset containing both text and image data, extracting features using Term Frequency-Inverse Document Frequency (TF-IDF) for text analysis and the color correlogram and Gray-Level Co-occurrence Matrix (GLCM) for image analysis. To enhance classification accuracy, the authors employed a metaheuristic optimization algorithm called the Fitness Oriented Levy Improvement-based Dragonfly Algorithm (FLI-DA) for feature selection. The detection process combined recurrent neural networks (RNNs) and convolutional neural networks (CNNs), with FLI-DA further optimizing the number of hidden neurons. Experimental evaluations demonstrated significant improvements in spam email classification performance, reinforcing the effectiveness of hybrid deep learning approaches in phishing detection.

Further advancing phishing detection in the email domain, the study in [58] highlighted the growing vulnerability of organizations to phishing and spam emails, emphasizing the need for efficient detection strategies. To this end, the authors introduced Phish Responder, a hybrid machine learning framework that integrates long-short-term memory (LSTM) networks for text-based datasets and Multilayer Perceptron (MLP) for numerical-based datasets in natural language processing. Their experimental evaluation demonstrated that Phish Responder achieved an impressive accuracy of 99% using the LSTM model and 94% with the MLP model. Comparisons with existing solutions revealed the superiority of their approach, particularly in numerical-based phishing detection, thereby showcasing the potential of hybrid machine learning models in combating sophisticated phishing schemes. Based on the need for improved phishing detection systems, [59] developed a hybrid phishing detection model that combined URL and content analysis based on deep learning. The study

addressed the growing prevalence of phishing attacks, particularly those targeting victims through email communications. The proposed system aimed to improve detection accuracy while minimizing false positives by incorporating both URL-based and content-based features. Experimental evaluations, conducted on a high-risk URL and content-based phishing detection data set sourced from PhishTank, demonstrated a significant improvement in accuracy, reaching 98.37%. These results further highlight the advantages of integrating multiple detection mechanisms to enhance the accuracy of phishing identification while reducing erroneous classifications. The study in [60] explored the increasing threat of phishing attacks and proposed a deep learning-based detection model to improve the identification of phishing. By providing a comprehensive overview of phishing techniques, the authors emphasized the urgent need for more effective detection technologies. Their approach involved training and validating a model using split dataset, that analyzed email text and other relevant features to classify emails as phishing or non-phishing. The results demonstrated that the boosted decision tree algorithm achieved impressive accuracy rates of 88%, 100%, and 97% across different datasets, reinforcing the robustness and adaptability of the model to various phishing patterns. Expanding on ensemble learning strategies, the work in [61] introduced HELPHED, a phishing email detection framework that integrates hybrid features with Ensemble Learning methods. The methodology utilized two distinct approaches: Stacking Ensemble Learning and Soft Voting Ensemble Learning, each using multiple machine learning algorithms to improve classification accuracy. A rigorous evaluation process, which incorporates innovative guidelines and extensive experimentation, validated the effectiveness of the model. Conducted on a highly imbalanced dataset comprising 32,051 benign and 3,460 phishing email samples, the study revealed that the Soft Voting Ensemble Learning approach outperformed competing deep learning and machine learning models, achieving an outstanding F1 score of 0.9942. These findings highlight the potential of hybrid ensemble techniques in improving phishing detection. Similarly, [62] introduced the LBPS model, a hybrid deep neural network designed for the detection of phishing scam accounts. Using LSTM-FCN and BP neural networks, the model analyzed transaction records to establish implicit relationships between the extracted features. Additionally, the LSTM-FCN network effectively captured temporal patterns from all transaction records associated with a target account. The experimental evaluation demonstrated the effectiveness of the selected features, as the LBPS model consistently outperformed existing detection methods, achieving a remarkable F1-score of 97.86% on Ethereum transaction datasets. This research underscores the importance of hybrid deep learning approaches in identifying fraudulent accounts with high accuracy. Further advancing phishing detection methodologies, the study in [63] introduced Hybrid Feature Detection based on Hybrid Features (PDHF), a framework that integrates optimal artificial and automatic deep learning features. Artificial phishing features were optimized by eliminating redundancies using a novel evaluation index and an enhanced bidirectional search algorithm, while deep features were extracted from URLs through a one-dimensional CNN and a disorderly quantized attention mechanism. The experimental findings confirmed the superiority of PDHF's over existing phishing detection methods, achieving an exceptional accuracy of 99.65%, along with a precision of 99.42%, recall of 99.40%, and an F1-score of 99.41%. These results further validate the advantages of hybrid feature-based approaches in refining phishing detection accuracy. Building on these advances, [64] introduced RNT, a deep learning-based phishing detection technique that combines the ResNeXt architecture with an embedded GRU model. Addressing the class imbalance issues, the study incorporated the SMOTE technique, while autoencoders and ResNet (EARN) were used for enhanced feature engineering. Furthermore, hyperparameter optimization was performed using the Jaya optimization method (RNT-J), ensuring improved model performance. Experimental results demonstrated that RNT consistently outperformed state-of-the-art phishing detection models, achieving an impressive accuracy of 98% while maintaining low false positive and false negative rates. These findings highlight the effectiveness of combining deep learning architectures with optimized feature engineering to develop high-precision real-time phishing detection solutions.

In summary, this section highlights the widespread and continuously evolving nature of phishing attacks, which pose significant threats to individuals and organizations worldwide. Over the years, various detection methods have been proposed, including List-Based, Visual Similarity, Heuristic, Machine Learning, and Deep Learning techniques, each offering distinct advantages in identifying phishing websites. However, despite their effectiveness, most of these approaches rely on batch or offline machine learning models, limiting their adaptability to emerging and dynamic phishing strategies. Numerous studies have evaluated machine learning algorithms such as Random Forest, Support Vector Machine, Naïve Bayes, and Decision Tree across diverse datasets, demonstrating high accuracy in phishing detection. Furthermore, deep learning techniques, including Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN), have shown significant promise in improving detection capabilities through automated feature learning. Although these advances have contributed to improved accuracy, they remain largely dependent on static training models that require frequent retraining to accommodate new threats. Despite these progresses, a critical research gap persists in the development and application of incremental machine learning techniques specifically tailored for phishing detection. Given the dynamic and continuously evolving nature of phishing attacks, detection systems must incorporate adaptive learning mechanisms capable of updating models in real time. Addressing this limitation is crucial for enhancing the robustness and efficiency of phishing detection systems, ultimately leading to more resilient cybersecurity solutions. Hence, further research should focus on integrating incremental learning methodologies to ensure adaptability and sustained performance in real-world phishing detection scenarios.

III. Methodology

This section describes the key procedures implemented to evaluate the effectiveness of the selected incremental machine learning algorithms using the phishing URL dataset provided in [65], [66]. The evaluation process, illustrated in Fig. 1, is carried out using the River Incremental Machine Learning Platform [67]. Each incremental machine learning algorithm is assessed following the same evaluation procedure to ensure consistency and comparability. As depicted in Fig. 1, the evaluation process begins with loading the phishing URL

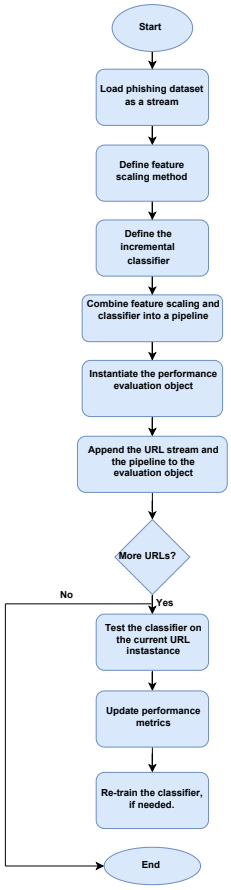


Fig. 1. Evaluation flowchart.

dataset as a continuous data stream in the River environment. Next, an incremental machine learning pipeline is constructed that incorporates a feature scaling method along with an incremental classifier. Once the pipeline is established, the phishing URL stream and the configured pipeline are integrated into the performance evaluation module, preparing the system for training and assessment. This setup ensures that the evaluation process accurately reflects the real-time adaptability of incremental learning models in phishing detection.

In this study, the performance of incremental machine learning classification algorithms is evaluated using the prequential evaluation method, also known as the interleaved test-then-train approach. This evaluation technique is specifically designed for incremental machine learning scenarios, where each incoming data sample, called an instance, plays a dual role in both testing and training the model. As instances are processed sequentially in their order of arrival, they become inaccessible after processing, ensuring that the model continuously adapts to new data without retaining past instances. In the context of phishing URL detection, the prequential evaluation process begins by using each incoming URL sample to test the incremental machine learning algorithm, generating a prediction. Immediately afterward, the same sample is utilized for training, allowing the model to refine its learning dynamically. Performance metrics are incrementally updated with each observed instance, allowing for a continuous assessment of the effectiveness of the model and its adaptability to previously unseen data. This ensures that the incremental machine learning algorithm undergoes a real-time evaluation, with its performance metrics continuously reflecting its ability to generalize to new phishing attempts. To comprehensively assess the effectiveness of the algorithms tested, several widely recognized performance metrics are used. These include accuracy, precision, recall, F score, kappa statistic, and memory footprint of the machine learning model during online learning. In the context of phishing URL detection, these metrics are formally defined as follows:

- Classification Accuracy: Represents the percentage of correctly classified URL samples [see Eq. (1)].

$$Accuracy = \frac{TN + TP}{TP + FP + FN + TN} \times 100\% \quad (1)$$

where, TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. TP represents the number of correctly classified phishing URLs, while TN corresponds to correctly classified benign URLs. In contrast, FP denotes benign URLs misclassified as phishing, whereas FN refers to phishing URLs misclassified as benign.

- Precision: Measures the proportion of URLs classified as phishing that actually belong to the phishing class [see Eq. (2)].

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

- Recall: Evaluates the proportion of phishing URLs correctly identified out of all phishing URLs in the observed stream [see Eq. (3)].

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

- F-score: Computes the harmonic mean of precision and recall, providing a balanced measure of a model's classification performance [see Eq. (4)].

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

- Cohen's Kappa (κ): A robust measure of classification accuracy that accounts for agreement by chance. Unlike traditional accuracy metrics, Cohen's Kappa highlights improvements over a majority-class classifier that predicts all instances as the most frequent class [68]. This is particularly useful when evaluating the classification performance in unbalanced data streams [see Eq. (5)].

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \quad (5)$$

where, p_0 represents the prequential accuracy of the classifier, while p_c denotes the probability that a chance classifier

correctly predicts labels [69]. A value of $\kappa = 1$ indicates perfect classification performance.

- Model Size: Represents the memory space occupied by the incremental machine learning model during the prequential evaluation process, reflecting its computational efficiency.

To comprehensively evaluate incremental machine learning algorithms, an extensive set of experiments was conducted using various models, as detailed in Table I. This study explores a diverse range of incremental machine learning approaches, including ensemble-based methods, forest models, linear classifiers, Naïve Bayes, nearest neighbor models, and tree-based classifiers. Additionally, each algorithm was tested under three different feature scaling settings: no scaling, min-max scaling, and standard scaling.

The Min-Max scaling, defined in Eq. (6), transforms the data to a fixed range between 0 and 1, preserving the relative features relationships. In contrast, standard scaling, as described in Eq. (7), adjusts the data to have zero mean and unit variance, ensuring that the features contribute equally to model training. In this study, the impact of these scaling techniques on the performance of incremental learning models is systematically analyzed.

TABLE I. UTILIZED INCREMENTAL MACHINE LEARNING ALGORITHMS

Category	Algorithm Name
Ensemble Models	ADWINBagging Classifier ADWINBoosting Classifier AdaBoost Classifier BOLE Classifier Bagging Classifier Leveraging Bagging Classifier Streaming Random Patches (SRP) Classifier Stacking Classifier Voting Classifier
Forest Models	Aggregated Mondrian Forest (AMF) Classifier Adaptive Random Forest (ARF) Classifier
Linear Models	Logistic Regression (LR) Classifier Approximate Large Margin Algorithm (ALMA) Classifier Passive-aggressive (PA) Classifier Perceptron Classifier
Naïve Bayes Models	Gaussian (NB) Classifier
Neighbours Models	K-Nearest Neighbors (KNN) Classifier
Tree Models	Hoeffding Tree (HT) Classifier Hoeffding Adaptive Tree (HAT) Classifier Extremely Fast Decision Tree (EFDT) Classifier Stochastic Gradient Tree (SGT) Classifier

$$x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{6}$$

where,

- x denotes the original feature value.
- x_{scaled} represents the scaled feature value.
- x_{min} refers to the running minimum value of the feature in the data stream.
- x_{max} refers to the maximum running value of the feature in the data stream.

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \tag{7}$$

where,

- x represents the original feature value.
- x_{scaled} denotes the scaled value of the feature.
- μ corresponds to the running mean of the feature in the data stream.
- σ denotes the running standard deviation of the feature in the data stream.

IV. RESULTS

This section presents the results obtained by implementing the methodology described in Fig. 1 on the incremental machine learning algorithms summarized in Table I, evaluated using different feature scaling techniques. The objective is to analyze the impact of feature scaling on model performance and memory efficiency.

Fig. 2 illustrate the performance results for linear models in standard, min-max, and no feature scaling settings, respectively. The results show that the scaling of the features plays a crucial role in influencing the accuracy, precision, recall, F1 score, Cohen's kappa and model size.

As shown, the application of standard feature scaling significantly enhances the performance of the logistic regression model, which outperforms its linear counterparts. Specifically, logistic regression achieves an accuracy of 89.20%, with precision, recall, F1 score, and Cohen's kappa values of 86.42%, 89.42%, 87.89% and 78.15%, respectively. This improvement suggests that standard scaling effectively normalizes feature distributions, allowing the model to learn more efficiently and generalize better. On the other hand, the ALMA (Adaptive Linear Margin Algorithm) model exhibits the lowest memory consumption, occupying only 4.69 KB, making it the most resource-efficient among the linear models. This result highlights a critical trade-off between accuracy and memory efficiency, where ALMA prioritizes compactness at the cost of predictive performance.

A different performance pattern emerges under Min-Max scaling. The Perceptron model achieves the highest accuracy of 83.60%, indicating that the min-max scaling may provide a more favorable feature distribution for this model. Meanwhile, logistic regression maintains the highest precision value of 90.70%, suggesting that min-max scaling helps reduce false positive classifications in phishing detection. The Passive-Aggressive (PA) model, in contrast, attains the highest recall (81.57%), F1 score (81.27%), and Cohen's kappa (66.56%), indicating that it is particularly adept at capturing phishing instances under this scaling technique. Consistent with the previous setting, the ALMA model retains the lowest memory footprint, occupying only 6.56 KB.

When no feature scaling is applied, as shown in Fig. 2, the performance of the model remains relatively similar to the Min-Max scaling scenario, but with slight variations in key metrics. The Perceptron model continues to exhibit the highest accuracy (83.60%), while logistic regression achieves the highest precision (91.05%). The PA model, once again, records the highest recall (81.57%), F1 score (81.27%), and Cohen's kappa (66.56%), reinforcing its strong capability to identify phishing instances. In particular, the ALMA model

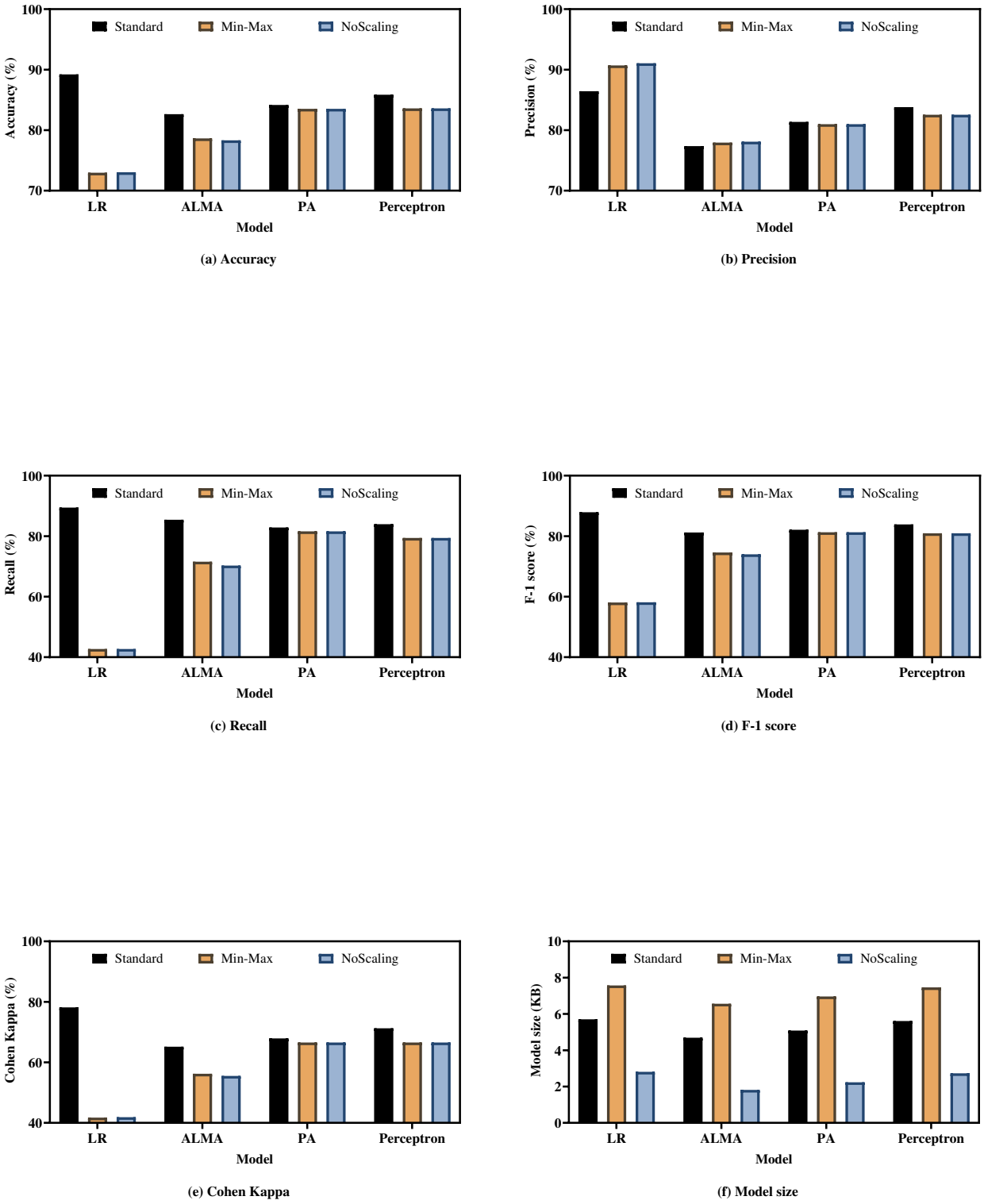


Fig. 2. Linear models' performance.

maintains its efficiency in terms of memory consumption, with a significantly reduced model size of just 1.81 KB, demonstrating its adaptability to unscaled feature spaces.

In general, these findings highlight the substantial impact of feature scaling on the performance of linear incremental learning models. Standard scaling generally enhances classification accuracy and stability, particularly benefiting logistic regression, while min-max scaling provides a more favorable

environment for models such as the Perceptron and PA. The absence of scaling appears to produce results comparable to min-max scaling, though with minor fluctuations in performance metrics. Additionally, ALMA consistently maintains the lowest memory usage in all settings, making it the most suitable choice for resource-constrained environments. These results emphasize the importance of selecting appropriate feature scaling techniques based on the specific trade-offs between accuracy and computational efficiency required for phishing detection tasks.

Fig. 3 illustrates the performance metrics obtained for the incremental machine learning models of Naive Bayes (NB) and K-Nearest Neighbors (KNN) under different feature scaling techniques. Interestingly, both NB and KNN models demonstrate a high degree of consistency in their performance across all feature scaling settings, indicating their relative robustness to variations in feature distributions.

As depicted, the KNN model achieves superior performance in terms of accuracy, precision, F1-score, and Cohen's kappa, attaining values of 88.87%, 86.04%, 87.51% and 77.48%, respectively. These results suggest that KNN benefits from its ability to classify new data points based on proximity in the feature space, which remains stable even after standard scaling is applied. In contrast, the NB model excels in recall, achieving a value of 89.40%, indicating its strength to correctly identify phishing instances. Additionally, the NB model maintains a smaller memory footprint, occupying 61.45 KB, which is notably more efficient compared to KNN.

This performance pattern remains largely consistent under min-max and no-scaling cases, where NB and KNN models continue to exhibit comparable trends, albeit with slight variations in individual performance metrics. The persistence of these trends across different scaling methods suggests that both NB and KNN models inherently adapt well to diverse feature distributions, making them relatively insensitive to the choice of feature scaling.

In general, these findings highlight the resilience of the NB and KNN models to feature scaling, maintaining consistent classification performance regardless of the applied normalization technique. This robustness makes them suitable candidates for phishing detection tasks, where feature distributions may vary dynamically. Furthermore, the trade-off between model size and performance suggests that, while KNN offers higher predictive accuracy, NB remains a computationally efficient alternative, making it particularly useful in scenarios with resource constraints.

Fig. 4 illustrates the performance of several tree-based models under different feature scaling techniques. The results indicate that tree-based models demonstrate relatively stable performance across various scaling settings, with minimal variations in classification accuracy and other key metrics.

As shown, under standard feature scaling, the Extremely Fast Decision Tree (EFDT) emerges as the best performing model, achieving the highest accuracy (88.47%), recall (88.85%), F1-score (87.10%) and Cohen's kappa (76.68%). These results suggest that EFDT effectively learns from streaming data while maintaining strong generalization capabilities. In contrast, the Hoeffding Tree (HT) algorithm attains the highest precision (85.51%) and maintains the lowest memory footprint, with a model size of 139.03 KB, demonstrating its computational efficiency.

Moving to Min-Max scaling, EFDT continues to achieve the highest accuracy (88.63%), F1-score (87.16%) and Cohen's kappa (76.96%), reinforcing its effectiveness across different scaling techniques. However, performance differences become more pronounced among other tree-based models. The Hoeffding Adaptive Tree (HAT) attains the highest precision (87.10%), while the Stochastic Gradient Tree (SGT) model achieves the highest recall (88.32%). Despite these variations, the HT algorithm retains its advantage in memory efficiency, maintaining a compact model size of 140.81 KB.

When no feature scaling is applied, EFDT continues to lead in accuracy (88.79%), F1-score (87.34%), and Cohen's kappa (77.29%), highlighting its robustness in handling raw feature distributions. Meanwhile, the HT model exhibits the highest precision (87.52%), further supporting its tendency to reduce false positives. Furthermore, the SGT model records the highest recall (88.32%), strengthening its ability to detect phishing instances effectively. The HT model maintains its efficiency in memory consumption, achieving the smallest model size of 131.73 KB.

In general, tree-based models exhibit minimal performance variations in different feature scaling settings, suggesting their inherent ability to adapt to varying feature distributions without requiring extensive normalization. The EFDT model consistently outperforms its counterparts in accuracy and generalization metrics, making it a strong candidate for phishing detection tasks. Meanwhile, the HT model remains the most memory-efficient, balancing classification performance with low computational overhead. These findings suggest that, while feature scaling can have a significant impact on certain machine learning models, tree-based algorithms remain relatively resilient, ensuring consistent performance across different preprocessing techniques.

Fig. 5 illustrates the performance of the Aggregated Mondrian Forest (AMF) and Adaptive Random Forest (ARF) models under different feature scaling techniques. The results indicate that both AMF and ARF exhibit consistently strong classification performance across all scaling settings, highlighting their robustness in handling varying feature distributions.

As shown, under standard feature scaling, the AMF model achieves the highest performance across multiple metrics, including accuracy (90.15%), precision (88.83%), F1-score (88.75%), Cohen's kappa (79.99%), and model size (3051.52 KB). These results suggest that AMF effectively leverages feature standardization to enhance classification performance while maintaining a relatively compact memory footprint. Meanwhile, the ARF model demonstrates a marginally higher recall value of 88.85%, indicating its strength in correctly identifying phishing instances while slightly compromising precision. A similar performance trend is observed when the Min-Max feature scaling is applied. Both AMF and ARF models continue to maintain high classification performance, with minimal deviations in accuracy, precision, and recall. This consistency reinforces their adaptability to different feature normalization techniques, making them well-suited for real-world phishing detection scenarios.

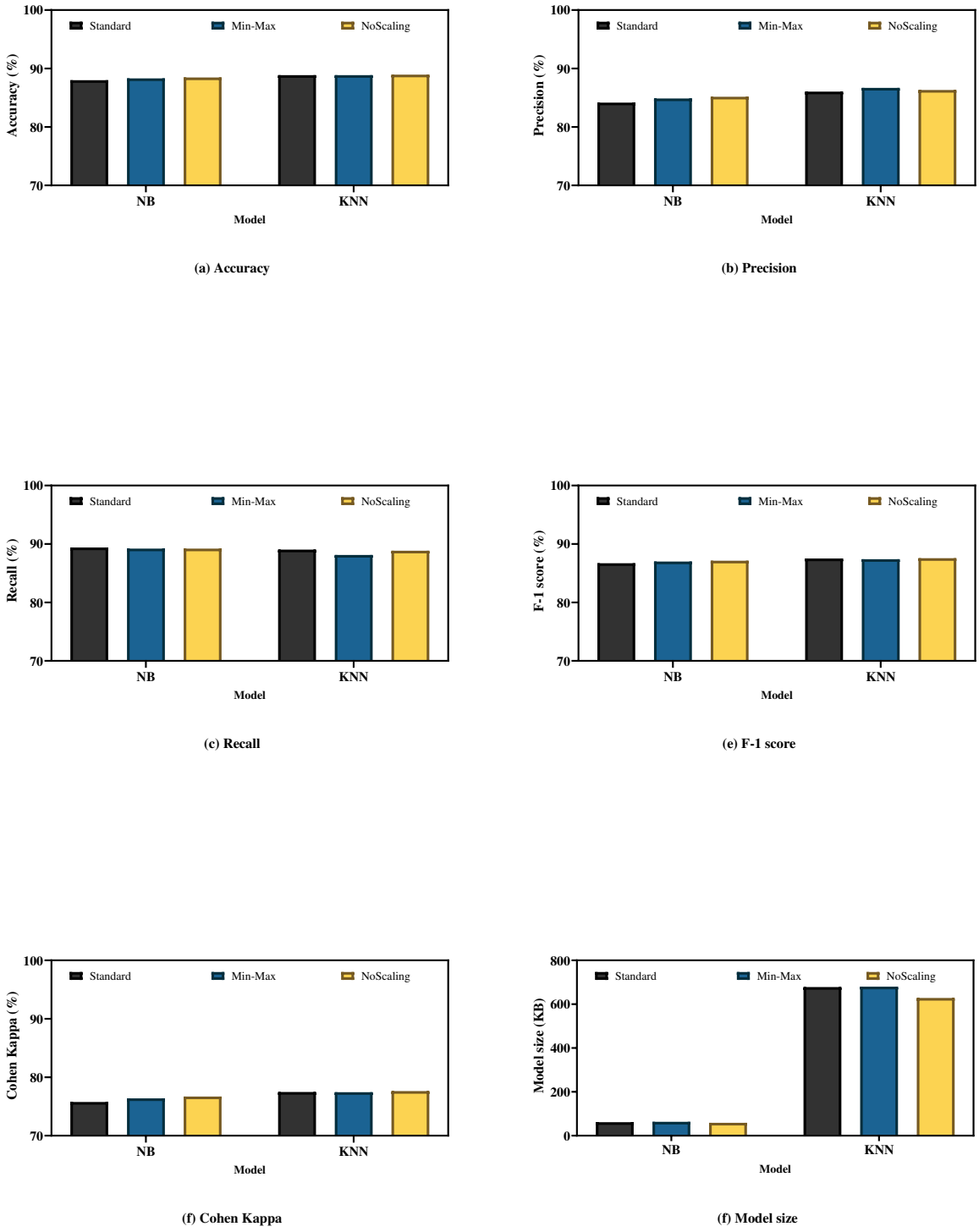


Fig. 3. Naive and Neighbour's models.

When no feature scaling is applied, both models experience a slight decrease in classification performance. However, in this setting, ARF marginally outperforms AMF, achieving an accuracy of 89.83%, precision of 88.04%, recall of 88.85%, F1-score of 88.44%, and Cohen's kappa of 79.37%. Despite this, the AMF model maintains an advantage in memory efficiency, requiring only 3000.32 KB compared to the larger footprint of the ARF. This trade-off suggests that while ARF can slightly improve predictive performance in unscaled data, AMF remains the more memory-efficient option.

In general, these findings indicate that AMF and ARF mod-

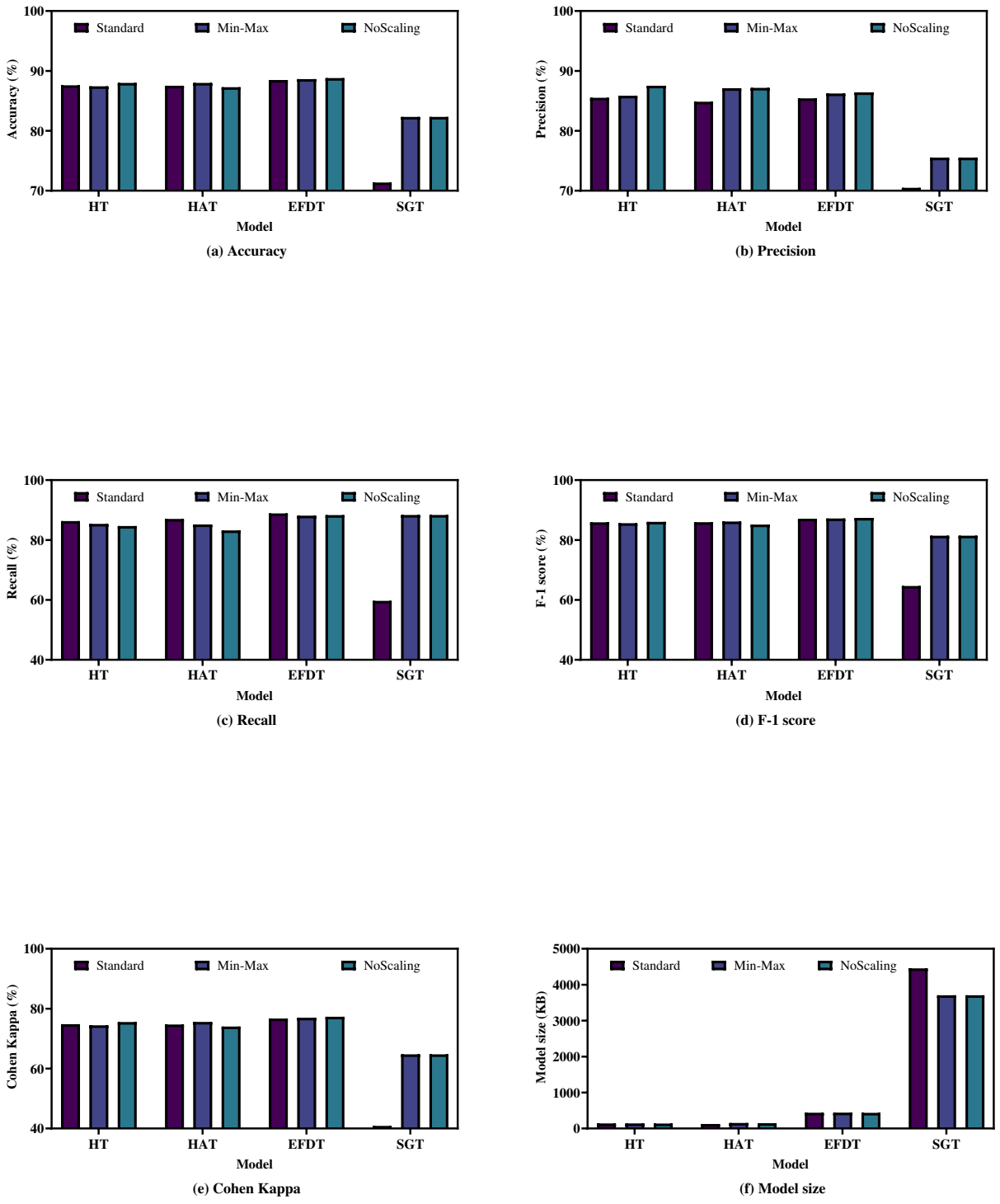


Fig. 4. Tree-based models' performance.

els exhibit resilience to feature scaling techniques, consistently achieving high classification accuracy regardless of the applied normalization approach. The AMF model excels in terms of memory efficiency while maintaining strong classification performance, while the ARF model demonstrates slightly superior recall and generalization capabilities. These insights suggest that both models are well suited for phishing detection tasks, particularly in environments with varying data preprocessing constraints.

Fig. 6 presents the performance evaluation of ensemble-based incremental machine learning models under different

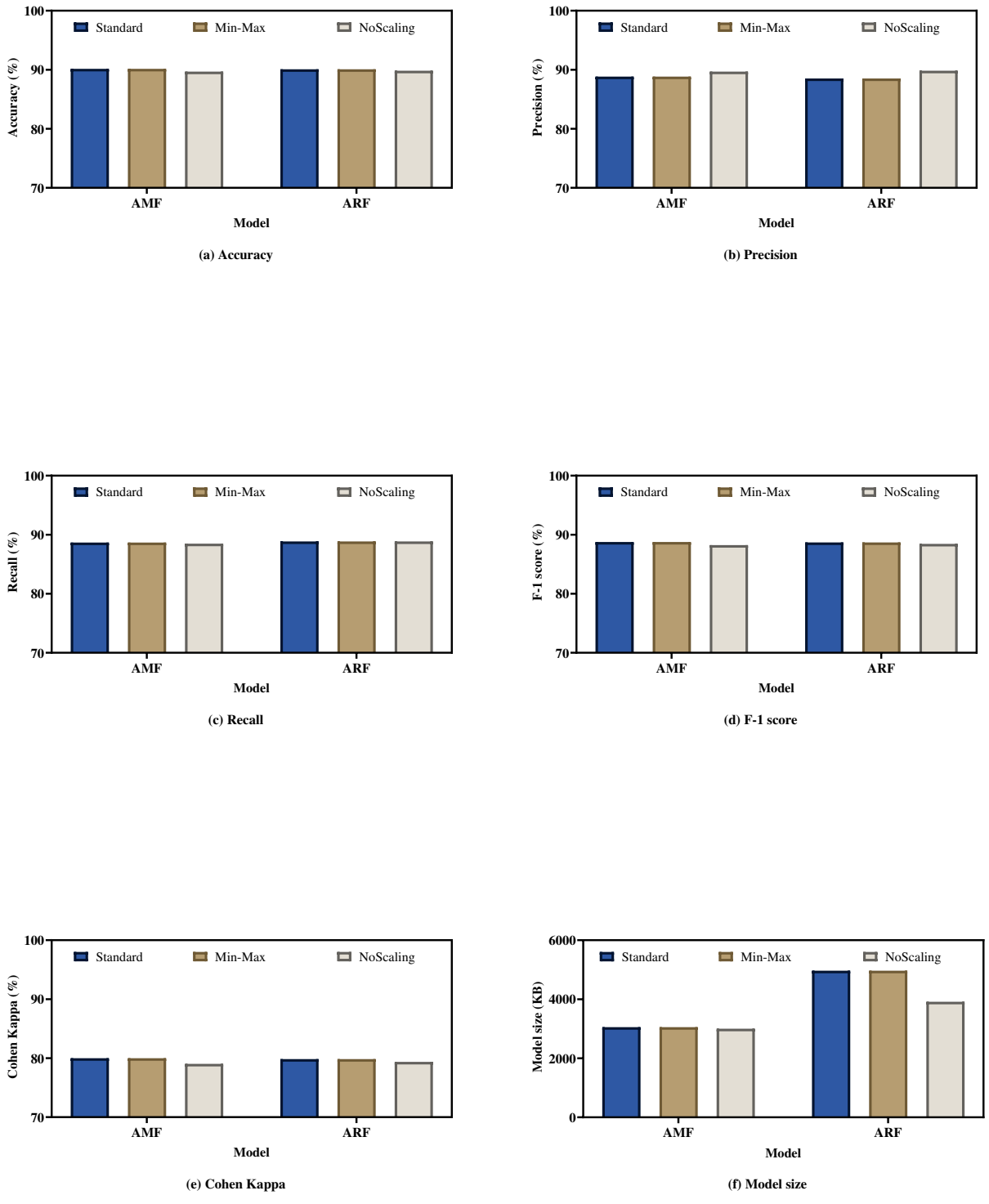


Fig. 5. Forest models' performance.

feature scaling techniques. Unlike other model categories, the ensemble methods exhibit noticeable variations in performance metrics depending on the applied feature scaling strategy.

As depicted, when standard feature scaling is applied, no single ensemble model dominates across all performance metrics. The Leveraging Bagging (LB) technique achieves the highest accuracy at 89.68%, demonstrating its robustness in classifying phishing URLs correctly. Meanwhile, the stacking ensemble outperforms other models in terms of precision, attaining a value of 88.38%, suggesting its effectiveness in

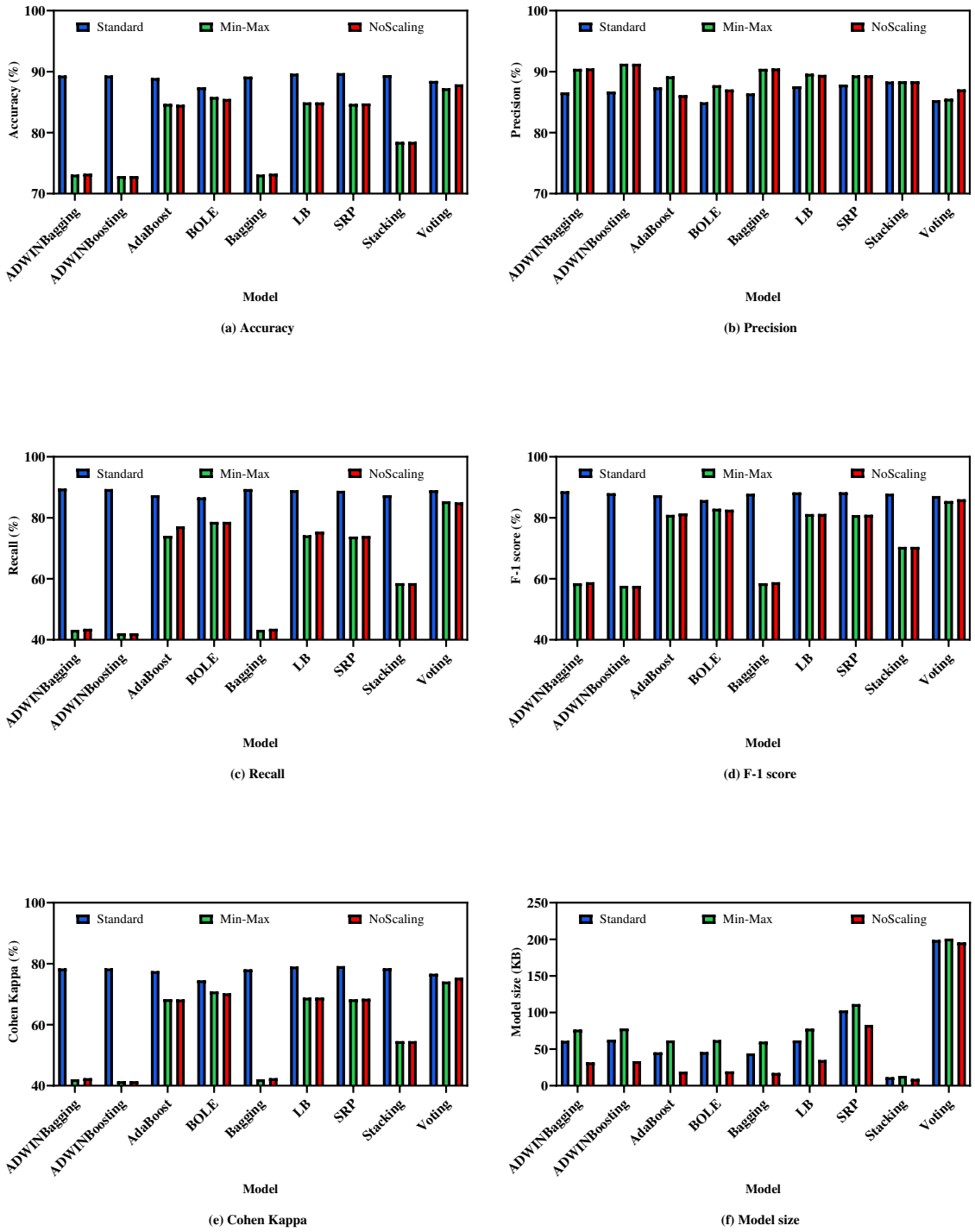


Fig. 6. Ensemble models' performance.

minimizing false positives. The ADWINBagging ensemble stands out in recall (89.60%) and F1-score (88.70%), indicating its strength in correctly identifying phishing instances. Additionally, the Streaming Random Patches (SRP) ensemble secures the highest Cohen's kappa value (79.21%), reflecting its strong overall classification reliability. Regarding model size, the Stacking ensemble maintains the smallest memory footprint, requiring only 11.55 KB, making it the most efficient in terms of storage.

On the other hand, when Min-Max feature scaling is applied, a decline in most performance metrics is observed, except for precision, which sees slight improvements. The Voting ensemble model achieves the highest accuracy (87.28%), recall

(85.40%), F1-score (85.48%), and Cohen's kappa (74.16%), indicating that it adapts well to min-max scaling despite the overall performance reduction. The ADWINBoosting ensemble exceeds others in precision, reaching 91.30%, suggesting that it maintains strong decision boundaries even when the feature values are normalized to a fixed range. The Stacking ensemble continues to be the most memory-efficient, requiring 13.43 KB of storage.

A similar trend is observed in case of no-scaling, which presents the performance of ensemble models when no feature scaling is applied. The Voting ensemble maintains its advantage in accuracy (87.92%), recall (85.04%), F1-score (86.06%), and Cohen's kappa (75.40%), reinforcing its stability across different feature preprocessing conditions. The ADWINBoosting model, once again, records the highest precision (91.30%), indicating its consistent ability to minimize false positives. The Stacking ensemble remains the most compact model, occupying only 9.35 KB of memory, further highlighting its efficiency in resource-constrained environments.

In general, these results underscore the sensitivity of ensemble-based models to feature scaling techniques. While certain models, such as Voting and ADWINBoosting, maintain strong classification performance across multiple scaling settings, others, such as Stacking, optimize for memory efficiency. The variation in performance metrics suggests that selecting an appropriate ensemble method should be guided by the specific trade-offs between accuracy, precision, and computational efficiency required for phishing detection tasks.

V. Conclusion

This study has provided a comprehensive examination of the evolving cybersecurity landscape, particularly in the context of phishing detection. By systematically analyzing incremental machine learning methods and online model selection strategies leveraging multi-armed bandits, valuable insights have been gained that have broad implications for enhancing cybersecurity defenses.

The findings underscore the intricate performance variations exhibited by different machine learning models under various scaling conditions of features. From linear models to tree-based and ensemble methods, each algorithm demonstrated distinct sensitivities to feature scaling adjustments, emphasizing the necessity of tailoring detection methodologies to specific operational contexts. In particular, while some models exhibited resilience to scaling changes in the feature, others displayed significant performance fluctuations, highlighting the critical role of pre-processing techniques in optimizing detection accuracy.

In general, this study contributes to the broader discourse on cybersecurity by offering practical insights into optimizing incremental machine learning approaches and evaluating model selection driven by reinforcement learning for phishing detection. By elucidating the nuanced performance characteristics of various models and selection strategies, this work lays the foundation for the development of more resilient, adaptive, and intelligent cybersecurity frameworks. Ultimately, these advances will strengthen real-time defenses against evolving cyber threats, reinforcing the robustness of modern cybersecurity systems.

Acknowledgment

Machine learning training and evaluation have been performed using the Phoenix High Performance Computing facility at the American University of the Middle East, Kuwait.

References

[1] R. Goenka, M. Chawla, and N. Tiwari, "A comprehensive survey of phishing: mediums, intended targets, attack and defence techniques and a novel taxonomy," *International Journal of Information Security*, vol. 23, no. 2, pp. 819–848, Apr 2024. [Online]. Available: https://doi.org/10.1007/s10207-023-00768-x

[2] T. Stojnic, D. Vatsalan, and N. A. G. Arachchilage, "Phishing email strategies: Understanding cybercriminals' strategies of crafting phishing emails," *SECURITY AND PRIVACY*, vol. 4, no. 5, p. e165, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.165

[3] K. Thakur, M. L. Ali, M. A. Obaidat, and A. Kamruzzaman, "A systematic review on deep-learning-based phishing email detection," *Electronics*, vol. 12, no. 21, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/21/4545

[4] L. Gallo, D. Gentile, S. Ruggiero, A. Botta, and G. Ventre, "The human factor in phishing: Collecting and analyzing user behavior when reading emails," *Computers and Security*, vol. 139, p. 103671, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404823005813

[5] R. Zieni, L. Massari, and M. C. Calzarossa, "Phishing or not phishing? a survey on the detection of phishing websites," *IEEE Access*, vol. 11, pp. 18 499–18 519, 2023.

[6] S. R. Borra, B. Gayathri, B. Rekha, B. Akshitha, and B. Hafeeza, "K-nearest neighbour classifier for url-based phishing detection mechanism," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 14, no. 03, pp. 34–40, 2023.

[7] C. Pires and J. Borges, "Detecting targeted phishing websites for brand protection and cyber defence using computer vision," in *2023 IEEE International Workshop on Technologies for Defense and Security (TechDefense)*, 2023, pp. 1–6.

[8] C. C. L. Tan, K. L. Chiew, K. S. Yong, Y. Sebastian, J. C. M. Than, and W. K. Tiong, "Hybrid phishing detection using joint visual and textual identity," *Expert Systems with Applications*, vol. 220, p. 119723, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423002245

[9] A. Garg, B. Sharma, A. Gupta, and R. Khan, "Security of modern networks and its challenges," in *Cyber Security Using Modern Technologies*. CRC Press, 2023, pp. 57–71.

[10] S. A. Afaq, M. S. Husain, A. Bello, and H. Sadia, "A critical analysis of cyber threats and their global impact," in *Computational Intelligent Security in Wireless Communications*. CRC Press, 2023, pp. 201–220.

[11] M. S. Kheruddin, M. A. E. M. Zuber, and M. M. M. Radzai, "Phishing attacks: Unraveling tactics, threats, and defenses in the cybersecurity landscape," *Authorea Preprints*, 2024.

[12] A. Karim, M. Shahroz, K. Mustofa, S. B. Belhaouari, and S. R. K. Joga, "Phishing detection system through hybrid machine learning based on url," *IEEE Access*, vol. 11, pp. 36 805–36 822, 2023.

[13] P. Pandey and N. Mishra, "Phish-sight: A new approach for phishing detection using dominant colors on web pages and machine learning," *International Journal of Information Security*, vol. 22, no. 4, pp. 881–891, 2023.

[14] V. Ganganwar, S. I. Hussain, A. Powar, S. Gaur, and A. Kumar, "Using website content for detecting phishing urls: A novel approach," in *Advances in Data-Driven Computing and Intelligent Systems*, S. Das, S. Saha, C. A. Coello Coello, and J. C. Bansal, Eds. Singapore: Springer Nature Singapore, 2024, pp. 87–102.

[15] L. Tang and Q. H. Mahmoud, "A survey of machine learning-based solutions for phishing website detection," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 672–694, 2021.

[16] E. Gandotra and D. Gupta, "An efficient approach for phishing detection using machine learning," *Multimedia Security: Algorithm Development, Analysis and Applications*, pp. 239–253, 2021.

[17] A. Almomani, M. Alauthman, M. T. Shatnawi, M. Alweshah, A. Alrosan, W. Alomoush, and B. B. Gupta, "Phishing website detection with semantic features based on machine learning classifiers: a comparative study," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 18, no. 1, pp. 1–24, 2022.

[18] Z. Alshingiti, R. Alaqel, J. Al-Muhtadi, Q. E. U. Haq, K. Saleem, and M. H. Faheem, "A deep learning-based phishing detection system using cnn, lstm, and lstm-cnn," *Electronics*, vol. 12, no. 1, p. 232, 2023.

[19] S. Jalil, M. Usman, and A. Fong, "Highly accurate phishing url detection based on machine learning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 7, pp. 9233–9251, 2023.

[20] S. Das Guptta, K. T. Shahriar, H. Alqahtani, D. Alsalman, and I. H. Sarker, "Modeling hybrid feature-based phishing websites detection using machine learning techniques," *Annals of Data Science*, vol. 11, no. 1, pp. 217–242, 2024.

[21] S. Pallaiyah, M. Amir, I. S. Mathew, S. S. Varma, S. Shakeer, M. Rajendran, and J. Govindaraj, "Url phishing detection using machine learning," in *AIP Conference Proceedings*, vol. 3042, no. 1. AIP Publishing, 2024.

[22] N. A. Azeez, S. Misra, I. A. Margaret, L. Fernandez-Sanz, and S. M. Abdulhamid, "Adopting automated whitelist approach for detecting phishing attacks," *Computers and Security*, vol. 108, p. 102328, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404821001528

[23] J. Bharadiya, "Machine learning in cybersecurity: Techniques and challenges," *European Journal of Technology*, vol. 7, no. 2, pp. 1–14, 2023.

[24] X. Li, Y. Zhou, Z. Jin, P. Yu, and S. Zhou, "A classification and novel class detection algorithm for concept drift data stream based on the cohesiveness and separation index of mahalanobis distance," *Journal of Electrical and Computer Engineering*, vol. 2020, p. 4027423, Mar 2020. [Online]. Available: https://doi.org/10.1155/2020/4027423

[25] L. Rutkowski, M. Jaworski, and P. Duda, *Basic Concepts of Data Stream Mining*. Cham: Springer International Publishing, 2020, pp. 13–33.

[26] M. Al-Tarawneh, "Data stream classification algorithms for workload orchestration in vehicular edge computing: A comparative evaluation," *The International Journal of Fuzzy Logic and Intelligent Systems*, vol. 21, no. 2, pp. 101–122, Jun 2021. [Online]. Available: http://www.ijfis.org/journal/view.html?doi=10.5391/IJFIS.2021.21.2.101

[27] U. Adhikari, T. H. Morris, and S. Pan, "Applying hoeffding adaptive trees for real-time cyber-power event and intrusion classification," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4049–4060, 2018.

[28] D. Nallaperuma, R. Nawaratne, T. Bandaragoda, A. Adikari, S. Nguyen, T. Kempitiya, D. De Silva, D. Alahakoon, and D. Pothuhera, "Online incremental machine learning platform for big data-driven smart traffic management," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4679–4690, 2019.

[29] J. S. Rojas, A. Pekar, A. Rendón, and J. C. Corrales, "Smart user consumption profiling: Incremental learning-based ott service degradation," *IEEE Access*, vol. 8, pp. 207 426–207 442, 2020.

[30] A. Alkhresheh, M. A. B. Al-Tarawneh, and M. Alnawayseh, "Evaluation of online machine learning algorithms for electricity theft detection in smart grids," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, 2022. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2022.0131096

[31] D. Jin, S. Chen, H. He, X. Jiang, S. Cheng, and J. Yang, "Federated incremental learning based evolvable intrusion detection system for zero-day attacks," *IEEE Network*, vol. 37, no. 1, pp. 125–132, 2023.

[32] A. GLAVAN and V. CROITORU, "Incremental learning for edge network intrusion detection," *REVUE ROUMAINE DES SCIENCES TECHNIQUES—SÉRIE ÉLECTROTECHNIQUE ET ÉNERGÉTIQUE*, vol. 68, no. 3, pp. 301–306, 2023.

[33] M. A. Shyaa, Z. Zainol, R. Abdullah, M. Anbar, L. Alzubaidi, and J. Santamaría, "Enhanced intrusion detection with data stream classification and concept drift guided by the incremental learning genetic programming combiner," *Sensors*, vol. 23, no. 7, p. 3736, 2023.

[34] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 590–611, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157823000034

[35] L. Yang, J. Zhang, X. Wang, Z. Li, Z. Li, and Y. He, "An improved elm-based and data preprocessing integrated approach for phishing detection considering comprehensive features," *Expert Systems with Applications*, vol. 165, p. 113863, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420306734

[36] A. K. Jain and B. B. Gupta, "Two-level authentication approach to protect from phishing attacks in real time," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 6, pp. 1783–1796, Nov 2018. [Online]. Available: https://doi.org/10.1007/s12652-017-0616-z

[37] M. H. Alkawaz, S. J. Steven, A. I. Hajamydeen, and R. Ramli, "A comprehensive survey on identification and analysis of phishing website based on machine learning methods," in *2021 IEEE 11th IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, 2021, pp. 82–87.

[38] H. F. Atlam and O. Oluwatimilehin, "Business email compromise phishing detection based on machine learning: A systematic literature review," *Electronics*, vol. 12, no. 1, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/1/42

[39] C. Catal, G. Giray, B. Tekinerdogan, S. Kumar, and S. Shukla, "Applications of deep learning for phishing detection: a systematic literature review," *Knowledge and Information Systems*, vol. 64, no. 6, pp. 1457–1500, Jun 2022. [Online]. Available: https://doi.org/10.1007/s10115-022-01672-x

[40] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from urls," *Expert Systems with Applications*, vol. 117, pp. 345–357, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418306067

[41] C.-Y. Wu, C.-C. Kuo, and C.-S. Yang, "A phishing detection system based on machine learning," in *2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA)*. IEEE, 2019, pp. 28–32.

[42] K. L. Chiew, C. L. Tan, K. Wong, K. S. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Information Sciences*, vol. 484, pp. 153–166, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025519300763

[43] M. T. Suleman and S. M. Awan, "Optimization of url-based phishing websites detection through genetic algorithms," *Automatic Control and Computer Sciences*, vol. 53, no. 4, pp. 333–341, Jul 2019. [Online]. Available: https://doi.org/10.3103/S0146411619040102

[44] H. Zuhair and A. Selamat, "Phishing hybrid feature-based classifier by using recursive features subset selection and machine learning algorithms," in *Recent Trends in Data Science and Soft Computing*, F. Saeed, N. Gazem, F. Mohammed, and A. Busalim, Eds. Cham: Springer International Publishing, 2019, pp. 267–277.

[45] J. Stobbs, B. Issac, and S. M. Jacob, "Phishing web page detection using optimised machine learning," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2020, pp. 483–490.

[46] M. Korkmaz, O. K. Sahingoz, and B. Diri, "Detection of phishing websites by using machine learning-based url analysis," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–7.

[47] G. Palaniappan, S. S, B. Rajendran, Sanjay, S. Goyal, and B. B S, "Malicious domain detection using machine learning on domain name features, host-based features and web-based features," *Procedia Computer Science*, vol. 171, pp. 654–661, 2020, third International Conference on Computing and Network Communications (CoCoNet'19). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920310383

[48] H. Shirazi, S. R. Muramudalige, I. Ray, and A. P. Jayasumana, "Improved phishing detection algorithms using adversarial autoencoder synthesized data," in *2020 IEEE 45th Conference on Local Computer Networks (LCN)*, 2020, pp. 24–32.

[49] W. Bai, "Phishing website detection based on machine learning algorithm," in *2020 International Conference on Computing and Data Science (CDS)*, 2020, pp. 293–298.

[50] A. Hannousse and S. Yahiouche, "Towards benchmark datasets for machine learning based website phishing detection: An experimental study," *Engineering Applications of Artificial*

Intelligence, vol. 104, p. 104347, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197621001950

[51] Ö. Kasim, "Automatic detection of phishing pages with event-based request processing, deep-hybrid feature extraction and light gradient boosted machine model," *Telecommunication Systems*, vol. 78, no. 1, pp. 103–115, Sep 2021. [Online]. Available: https://doi.org/10.1007/s11235-021-00799-6

[52] M. Sánchez-Paniagua, E. Fidalgo, V. González-Castro, and E. Alegre, "Impact of current phishing strategies in machine learning models for phishing detection," in *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, Á. Herrero, C. Cambra, D. Urda, J. Sedano, H. Quintián, and E. Corchado, Eds. Cham: Springer International Publishing, 2021, pp. 87–96.

[53] A. Butnaru, A. Mylonas, and N. Pitropakis, "Towards lightweight url-based phishing detection," *Future Internet*, vol. 13, no. 6, 2021. [Online]. Available: https://www.mdpi.com/1999-5903/13/6/154

[54] M. S. Munir Prince, A. Hasan, and F. Muhammad Shah, "A new ensemble model for phishing detection based on hybrid cumulative feature selection," in *2021 IEEE 11th IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, 2021, pp. 7–12.

[55] S.-J. Bu and S.-B. Cho, "Deep character-level anomaly detection based on a convolutional autoencoder for zero-day phishing url detection," *Electronics*, vol. 10, no. 12, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/12/1492

[56] M. Korkmaz, E. Kocyigit, O. K. Sahingoz, and B. Diri, "Phishing web page detection using n-gram features extracted from urls," in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2021, pp. 1–6.

[57] K. V. Samarthrao and V. M. Rohokale, "Enhancement of email spam detection using improved deep learning algorithms for cyber security," *Journal of Computer Security*, vol. 30, pp. 231–264, 2022, 2. [Online]. Available: https://doi.org/10.3233/JCS-200111

[58] M. Dewis and T. Viana, "Phish responder: A hybrid machine learning approach to detect phishing and spam emails," *Applied System Innovation*, vol. 5, no. 4, 2022. [Online]. Available: https://www.mdpi.com/2571-5577/5/4/73

[59] M. Korkmaz, E. Kocyigit, O. K. Sahingoz, and B. Diri, "A hybrid phishing detection system using deep learning-based url and content analysis," *Elektronika ir Elektrotechnika*, vol. 28, no. 5, pp. 80–89, Oct. 2022. [Online]. Available: https://eejournal.ktu.lt/index.php/elt/article/view/31197

[60] A. Mughaid, S. AlZu'bi, A. Hnaif, S. Taamneh, A. Alnajjar, and E. A. Elsoud, "An intelligent cyber security phishing detection system using deep learning techniques," *Cluster Computing*, vol. 25, no. 6, pp. 3819–3828, Dec 2022. [Online]. Available: https://doi.org/10.1007/s10586-022-03604-4

[61] P. Bountakas and C. Xenakis, "Helphed: Hybrid ensemble learning phishing email detection," *Journal of Network and Computer Applications*, vol. 210, p. 103545, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804522001862

[62] T. Wen, Y. Xiao, A. Wang, and H. Wang, "A novel hybrid feature fusion model for detecting phishing scam on ethereum using deep neural network," *Expert Systems with Applications*, vol. 211, p. 118463, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741742201555X

[63] E. Zhu, K. Cheng, Z. Zhang, and H. Wang, "Pdhf: Effective phishing detection model combining optimal artificial and automatic deep features," *Computers & Security*, vol. 136, p. 103561, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404823004716

[64] F. S. Alsubaei, A. A. Almazroi, and N. Ayub, "Enhancing phishing detection: A novel hybrid deep learning framework for cybercrime forensics," *IEEE Access*, vol. 12, pp. 8373–8389, 2024.

[65] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5948–5959, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417414001481

[66] "UCI Machine Learning Repository — archive.ics.uci.edu," http://archive.ics.uci.edu/dataset/379/website+phishing, [Accessed 03-04-2024].

[67] J. Montiel, M. Halford, S. M. Mastelini, G. Bolmier, R. Sourty, R. Vaysse, A. Zouitine, H. M. Gomes, J. Read, T. Abdessalem *et al.*, "River: machine learning for streaming data in python," 2021.

[68] T. Vasiloudis, F. Beligianni, and G. De Francisci Morales, "Boostvht: Boosting distributed streaming decision trees," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 899–908. [Online]. Available: https://doi.org/10.1145/3132847.3132974

[69] A. Bifet, G. de Francisci Morales, J. Read, G. Holmes, and B. Pfahringer, "Efficient online evaluation of big data stream classifiers," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 59–68. [Online]. Available: https://doi.org/10.1145/2783258.2783372

Automated Interactance Near-Infrared Spectral Acquisition System for Mandarin Quality Assessment

Van-Linh Lam[1], Dinh-Tri Nguyen[2], Thanh-Trung Le[3], Hoang-Tien Nguyen[4],
Phuoc-Loc Nguyen[5], Quoc-Khanh Huynh[6], Nhut-Thanh Tran[7], Chanh-Nghiem Nguyen[8]*
Faculty of Automation Engineering-College of Engineering, Can Tho University, Can Tho, Vietnam[1,2,3,4,7,8]
Automation Laboratory-College of Engineering, Can Tho University, Can Tho, Vietnam[1,2,3,4,8]
Faculty of Electric-Electronics, Kien Giang College, An Giang, Vietnam[1]
Faculty of Electric-Electronics and Computer, Kien Giang Vocational College, An Giang, Vietnam[5]
Faculty of Mechanical Engineering-College of Engineering, Can Tho University, Can Tho, Vietnam[6]

***Abstract*—Intelligent fruit grading systems require automated sensing solutions capable of rapid and reliable non-destructive internal quality assessment. While near-infrared spectroscopy has been widely used for soluble solids content (SSC) prediction, most existing studies rely on manually acquired spectra, limiting scalability in smart agricultural environments. Although online Vis–NIR systems based on transmission configurations have been reported, automated interactance-based systems designed for deployment-oriented grading remain limited. This study presents the design and validation of an automated interactance near-infrared spectral acquisition system for mandarin SSC evaluation. The system integrates controlled clamping, rotational positioning, and automated probe actuation to ensure stable optical geometry and repeatable probe–fruit contact during measurement. Spectral consistency was assessed by comparing consecutive scans obtained using manual and automated acquisition modes. The automated system reduced spectral dispersion among consecutive acquisitions within a measurement session by approximately 65% relative to manual measurement, indicating improved acquisition stability. Chemometric models based on partial least squares regression, support vector regression, and extremely randomized trees were developed under multiple preprocessing strategies. Prediction performance under automated acquisition remained within the same range as manual measurement, with several preprocessing–model combinations (particularly PLS and SVR with smoothing-based preprocessing), showing slightly higher R_{p} values and lower $\mathrm{RMSE}_{\mathrm{p}}$ values under automated acquisition. The findings demonstrate the feasibility of the proposed automated system for stable interactance spectral acquisition suitable for SSC prediction, supporting its potential future integration into automated fruit quality assessment systems.**

***Keywords*—Automated spectral acquisition; interactance spectroscopy; near-infrared spectroscopy; fruit quality assessment; soluble solids content; chemometric modeling; smart agriculture; sensing systems**

I. Introduction

Fruit quality evaluation plays a critical role in postharvest handling, grading, and commercialization of horticultural products. With the rapid development of smart agricultural technologies, there is increasing demand for automated sensing systems capable of delivering rapid, nondestructive, and reliable internal quality assessment under practical operational conditions. Among quality attributes, sweetness, commonly quantified by soluble solids content (SSC), is one of the most important indicators for mandarins. Conventional SSC measurement using refractometers requires juice extraction, making it destructive, time-consuming, and unsuitable for integration into automated grading systems [1]. Consequently, nondestructive optical sensing techniques have attracted increasing attention [2], particularly near-infrared (NIR) spectroscopy.

In NIR spectroscopy, incident radiation interacts with organic constituents within the fruit, inducing overtone and combination vibrations of C–H, N–H, and O–H bonds. The resulting spectral signatures contain information related to key quality attributes and can be effectively exploited using chemometric and machine-learning models for fruit quality prediction [3]. Depending on the optical configuration, spectral data can be acquired in reflectance, transmittance, or interactance modes. While reflectance measurements are influenced primarily by surface properties, transmittance measurements require high-intensity light sources to penetrate the entire sample. Interactance mode represents a practical compromise, capturing diffusely transmitted light that has penetrated into the fruit tissue, thereby enhancing sensitivity to internal quality attributes [4], [5].

Recent studies have developed online Vis–NIR systems for fruit quality assessment, mainly using transmission or reflectance configurations integrated with conveyor-based transport mechanisms [6], [7], [8], [9], [10]. For example, Wang et al. [10] reported an online diffuse transmission system for SSC detection in pomelo. Although such transmission systems enable continuous measurement, they usually require high-intensity light sources and are sensitive to fruit size, peel thickness, and internal structure. In contrast, interactance mode measures light that has penetrated into the fruit under controlled probe–fruit contact conditions, which can improve sensitivity to internal attributes while maintaining moderate optical penetration depth. However, automated interactance acquisition systems specifically designed to ensure controlled mechanical positioning and stable optical geometry for deployment-oriented grading applications remain relatively

*Corresponding author

limited.

For citrus fruits, peel composition and thickness significantly influence optical measurements. Although citrus peels may contain small amounts of free sugars, their contribution is relatively limited compared with the juice sac content [11], [12], suggesting that surface-dominated measurements may inadequately reflect internal sweetness. Interactance spectroscopy is therefore particularly suitable for mandarin SSC evaluation, as the porous and moderately thin peel permits partial penetration of NIR radiation into juice-containing tissues.

Despite extensive research on NIR-based fruit quality prediction, many reported systems remain laboratory-oriented and rely on manually operated spectral acquisition procedures. However, manual probe positioning can introduce variability associated with contact pressure, alignment differences, and operator handling, thereby limiting repeatability and scalability in automated grading environments. For practical smart agricultural deployment, sensing systems must ensure consistent optical conditions during measurement.

Preliminary feasibility experiments demonstrated that automated interactance spectral acquisition reduced the average measurement time per fruit from approximately 40 s to 17.7 s, corresponding to a 55.75% reduction and an estimated throughput of approximately 203 fruits per hour, while also improving measurement stability. At the same time, SSC prediction performance remained comparable to that of manual measurement, even under limited dataset size, preprocessing strategies, and regression models. These findings indicate that automation does not inherently degrade chemometric modeling performance and provide a foundation for further system-level development. However, these preliminary results remain limited to feasibility-level validation and do not yet fully address deployment-level requirements, including robustness, scalability, and operational reliability, for practical quality assessment systems. In addition, the study remains limited in terms of modeling configurations.

To address this gap, an automated interactance NIR spectral acquisition system is proposed. The system is an integrated sensing framework in which a dedicated acquisition module functions as the core measurement unit, coupled with data processing and modeling components for SSC prediction, thereby enabling integration into intelligent mandarin quality assessment systems. Within this framework, the acquisition module integrates controlled clamping, rotational positioning, and automated probe actuation to enable standardized, repeatable, and geometry-consistent spectral acquisition.

The proposed system emphasizes automated fruit positioning and controlled optical geometry to ensure repeatable spectral acquisition under operational conditions, thereby supporting reliable deployment in practical grading environments. Although the system was evaluated using mandarins in this study, the sensing architecture and acquisition strategy can be adapted to fruits with similar peel thickness and internal tissue structure, offering a scalable foundation for spectroscopy-based automated grading systems.

To evaluate whether automated acquisition preserves chemometric modeling performance, regression models including partial least squares regression, support vector regression, and extremely randomized trees were developed under multiple preprocessing strategies using a consistent cross-validation framework. Comparative analysis between manual and automated modes was conducted to determine whether observed differences arise from sensing configuration rather than modeling characteristics.

II. Methodology

A. Dataset Preparation

Fig. 1 illustrates the manual interactance spectral acquisition system. A halogen light source (MI-152, Dolan-Jenner) with a lamp power of 150 W was employed with sufficient intensity to penetrate the mandarin fruit, thereby satisfying the requirements for interactance spectral acquisition. Light from the source was delivered to the fruit through an MA4 ring-light optical fiber supplied by the manufacturer. The fiber output formed a circular illumination pattern that ensured uniform illumination at the measurement location. A liquid light guide (LLG5-4Z, Thorlabs) was used to transmit the interactance signal to the spectrometer (DLP NIRscan Nano EVM, Texas Instruments) which is capable of acquiring spectra in the wavelength range of 900–1700 nm. However, preliminary inspection indicated that signal intensities beyond approximately 1365 nm were extremely low ($< 3\%$ for automated acquisition and $\approx 0.6\%$ for manual acquisition), suggesting a noise-dominated region. Therefore, only the spectral range of 900–1365 nm was retained for subsequent analysis.

To enlarge the dataset, interactance spectra were collected from four measurement regions uniformly distributed along the equator of each mandarin. For each region, three adjacent measurement points were acquired and averaged to produce one representative spectrum used for modeling. A total of 460 mandarin fruits were used in this study. Because one representative spectrum was obtained from each of four measurement regions per fruit, the full dataset comprised 1840 averaged spectra for each acquisition mode (manual and automated). The spectral acquisition procedure for each region is illustrated in Fig. 2. Three measurement points were defined along the equatorial line of the fruit (Fig. 2a). A central point was first selected on the equator, followed by two additional points positioned 1 cm to the left and right of the central location. Spectra collected at the three measurement points were averaged to obtain a representative spectrum for the measurement region. This averaging strategy was adopted to reduce measurement noise and to mitigate the influence of spatial heterogeneity in SSC distribution.

B. Manual Acquisition of Interactance Spectra

At each measurement point, the mandarin was maintained in direct contact with the probe head (Fig. 2b) to block ambient light and ensure that only interactance light emerging from within the fruit was captured. This manual interactance spectral acquisition procedure is relatively time-consuming and is therefore mainly suitable for feasibility studies and laboratory-scale investigations. To overcome these limitations and enable high-throughput, non-destructive mandarin quality assessment, an automated interactance spectral acquisition module was developed, as described in the following section.

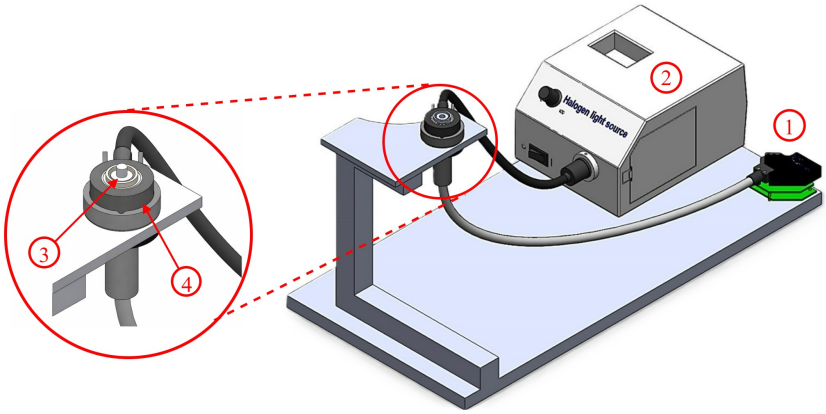
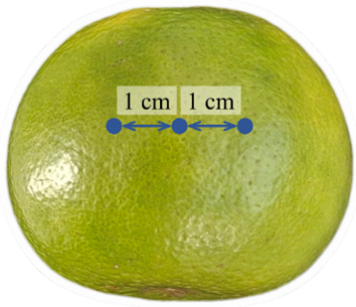
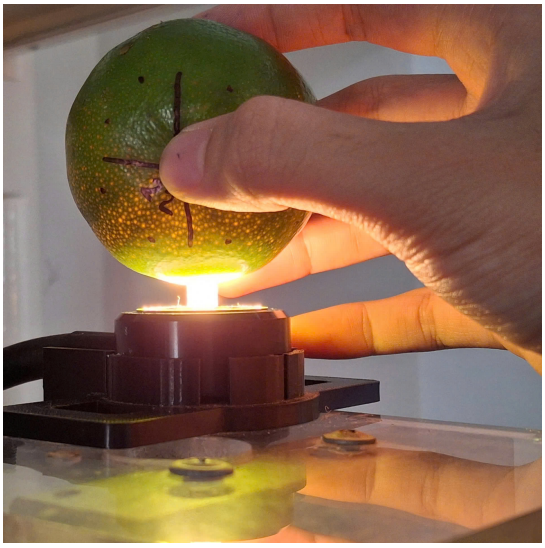


Fig. 1. Setup for collecting interactance spectra manually, including: (1) DLP NIRScan Nano EVM sensor, (2) MI-152 halogen light source, (3) the probe head, and (4) ring light output.



(a)



(b)

Fig. 2. (a) Spectral acquisition positions; (b) Interactance spectral measurement using the manual module.

C. Automated Interactance Spectral Acquisition

1) Overview of the system: The automated interactance spectral acquisition system provides controlled mechanical positioning and stable optical conditions for repeatable spectral measurements of mandarins. The system integrates fruit positioning, probe actuation, and optical isolation within a compact structure to enable automated multi-location spectral acquisition.

The overall mechanical configuration is shown in Fig. 3. The system is housed in a rigid frame forming a dark chamber that minimizes ambient light interference during measurement. Inside the chamber, a clamping and rotation mechanism positions the fruit, while an actuated spectral probe performs interactance measurements at predefined locations.

The clamping mechanism secures the mandarin between two curved contact heads to maintain stable positioning during measurement. Fruit presence and clamp position are monitored by sensors and limit switches to ensure consistent contact without excessive mechanical loading. The clamping frame is mounted on a rotary stage driven by a stepper motor, enabling controlled angular positioning for multi-location spectral acquisition.

The spectral probe–illumination unit is mounted on a horizontally actuated linear mechanism. During operation, the probe is automatically moved toward the fruit surface until a limit switch indicates contact, ensuring a repeatable probe–fruit interface. The probe integrates a ring-shaped illumination source surrounding the optical fiber used for spectral acquisition, enabling interactance measurements while reducing specular reflection.

To maintain measurement stability, the probe position, fruit orientation, and measurement sequence are automatically controlled by the system controller. This configuration enables repeatable spectral acquisition at multiple positions on the fruit surface while minimizing operator intervention.

2) Workflow of the system: Fig. 4 summarizes the operating workflow of the automated interactance spectral acquisition system. After fruit detection and clamping, the probe–illumination unit is advanced to acquire an interactance spectrum at a predefined measurement point, defined by a controlled angular position of the fruit. The fruit is then rotated sequentially to subsequent measurement points, and spectra are acquired following the same procedure. In this

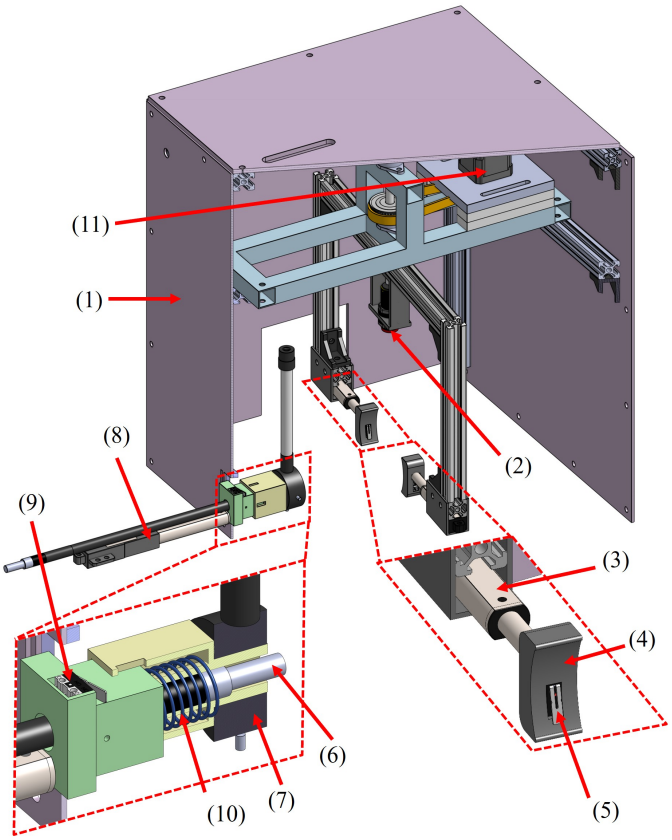


Fig. 3. Automated interactance spectral acquisition system and detailed views of its key mechanical components: (1) Dark chamber frame; (2) Proximity sensor for mandarin detection; (3) Clamp actuator; (4) Curved fruit contact head; (5) Integrated limit switch for contact detection; the acquisition module consisting of (6) Spectral probe head, (7) Ring-shaped illumination source, (8) Spectral probe–illumination unit actuator, (9) Limit switch for probe–fruit contact detection, (10) Compression spring for compliant probe contact; and (11) Stepper motor for rotating the clamp.

study, interactance spectra are collected at three circumferential measurement points per fruit to improve the spatial representativeness of the spectral measurement. Upon completion of the acquisition sequence, the probe retracts, the clamps release the fruit, and the probe returns to its initial state, ready for the next measurement cycle.

The automated interactance spectral acquisition system does not require manual predefinition of measurement positions. Based on a preliminary survey of the experimental samples, the average diameter of the sweet mandarins used for spectral acquisition was approximately 6.7 cm. Let $l = 1$ cm denote the arc length between two adjacent measurement points, $R \approx 3.35$ cm the average mandarin radius, and α (°) the required rotation angle of the fruit center to achieve an arc length of l. The rotation angle between two adjacent measurement points is calculated as:

$$\alpha = \frac{180\,l}{R\pi} = \frac{180 \times 1}{3.35\pi} \approx 17.1^\circ \tag{1}$$

The clamping frame is rotated via a belt transmission with a gear ratio $i = 1/2$ and driven by a stepper motor with a resolution of $N = 800$ steps per revolution. To rotate the

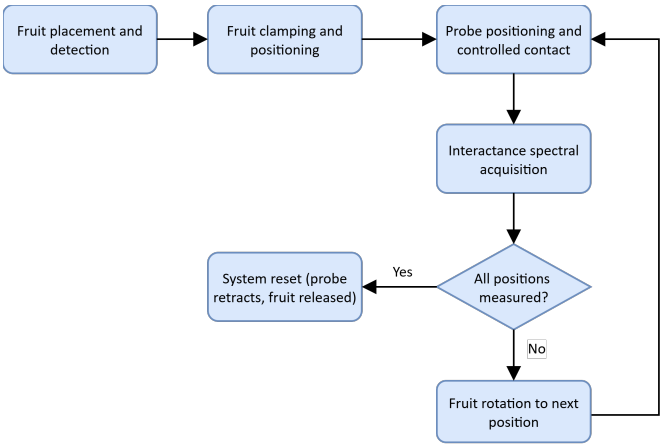


Fig. 4. Workflow of the automated interactance spectral acquisition system.

mandarin by an angle α, the required number of motor steps n is given by:

$$n = \frac{N\,\alpha}{360\,i} = \frac{800 \times 17.1}{360 \times 2} \approx 19 \text{ steps} \tag{2}$$

D. Performance Evaluation

1) Preliminary assessment of spectral consistency: To verify that the automated interactance spectral acquisition system provides spectra comparable in acquisition stability to those obtained using the manual setup, a preliminary assessment was conducted. This analysis focuses on evaluating the spectral consistency of consecutive acquisitions performed within a single measurement session, providing an initial indication of acquisition stability prior to regression model development. Improved spectral consistency is expected to enhance the reliability of the acquired data and is therefore likely to be reflected in the performance of subsequent chemometric models.

For each acquisition mode (manual and automated), four interactance spectra were acquired from consecutive acquisitions at the same nominal measurement region on the fruit surface under identical operating conditions within a single measurement session. Between consecutive acquisitions, the probe was repositioned according to the standard operating procedure of each mode. Consequently, measurements were performed within the same region but not at an identical physical contact point, thereby reflecting practical measurement variability associated with probe handling and positioning.

To quantify the consistency among repeated spectral measurements, the root mean square deviation (RMSD) was calculated as:

$$\text{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2} \tag{3}$$

where, n denotes the number of wavelengths in the spectrum and d_i is the standard deviation of the four repeated measurements at wavelength i, defined as:

$$d_i = \sqrt{\frac{1}{3}\sum_{j=1}^{4}(y_{ij} - \bar{y}_i)^2} \quad (4)$$

where, y_{ij} represents the interactance amplitude of the j-th replicate at wavelength i, and $\bar{y}_i$ is the mean amplitude at wavelength i calculated from the four repeated measurements. RMSD provides a wavelength-averaged indicator of spectral dispersion across the full spectral range and reflects variability among consecutive acquisitions.

A smaller RMSD indicates higher spectral consistency and, therefore, greater acquisition stability under the specified conditions. Because the measurements were conducted within a single session and confined to the same fruit region, this analysis primarily reflects instrumental and operational stability rather than broader sources of variability. For simplicity, inter-session variability across different days, operators, or fruit batches was not evaluated.

Therefore, this comparison should be interpreted as a screening-level assessment of acquisition stability. The overall feasibility of the automated acquisition system was further evaluated based on regression models calibrated using spectra acquired under each mode. Demonstrating reduced spectral dispersion under automated operation supports the reliability of the acquired data and confirms that improved spectral consistency does not adversely affect regression performance compared with manual acquisition. This result indicates that the automated acquisition system can provide spectral data of sufficient stability for subsequent regression model development for SSC prediction.

2) Regression model development: Regression models were developed to predict mandarin SSC (°Brix) from Vis–NIR interactance spectra acquired under both manual and automated acquisition modes. The primary objective was to evaluate whether spectra obtained using the automated interactance spectral acquisition system can support predictive models with performance comparable to those developed from manually acquired spectra.

As described in Section II-A, the study used 460 mandarin fruits, each measured at four equatorial regions, yielding 1840 averaged spectra per acquisition mode. The corresponding reference SSC values of these fruits ranged from 8.0 to 13.9 °Brix (mean 10.3 ± 1.14). For regression modeling, the dataset for each acquisition mode was partitioned into calibration (80%) and prediction (20%) subsets using the Sample set Partitioning based on joint X–Y distances (SPXY) algorithm [13], which selects representative samples by maximizing joint distances in both spectral (X) and response (SSC, Y) spaces. This strategy has been widely adopted to improve calibration representativeness and prediction reliability in NIR-based fruit quality modeling [14], [15], [16].

Prior to model calibration, spectral preprocessing was applied to reduce variability associated with scattering effects, baseline drift, and instrumental noise. The evaluated preprocessing methods included Min–Max scaling (MM), Standard Normal Variate (SNV), Savitzky–Golay (SG) smoothing, first-derivative (FD) transformation, and Mean–Median Absolute Deviation Normalization (MMADN), representing scaling, scatter correction, smoothing, derivative enhancement, and robust normalization strategies, respectively [17].

TABLE I. HYPERPARAMETER SEARCH RANGES AND FIXED SETTINGS FOR PLSR, SVR, AND ETR MODELS OPTIMIZED USING 10-FOLD CROSS-VALIDATION FOR MANDARIN SSC (BRIX) PREDICTION.

Model	Hyperparameter	Search range / fixed value
PLS	n_components	1 – N (number of wavelengths)
SVR	C	5 – 300 (step size: 5)
	ϵ (epsilon)	0.05–1.00 (step size: 0.05)
	cache_size	50, 100, 150
	Kernel	RBF
	Gamma	scale
ETR	n_estimators	10 – 200 (step = 10)
	max_depth	2 – 10 (step = 2)
	min_samples_split	default (2)
	min_samples_leaf	default (1)
	max_features	default (auto)
	bootstrap	False
	random_state	42
	criterion	squared_error

Because preprocessing selection can substantially influence model behavior and address different types of spectral artefacts [18], both single and combined preprocessing strategies were systematically evaluated to ensure a fair and unbiased comparison between manual and automated acquisition modes. Identical preprocessing procedures were applied to spectra from both acquisition modes to ensure comparability of regression results.

Three regression algorithms were investigated: Partial Least Squares regression (PLSR), Support Vector Regression (SVR), and Extremely Randomized Trees or Extra-Trees (ETR). PLSR was adopted as a baseline chemometric model, as it is described as a conventional regression method for SSC prediction in recent NIR fruit quality studies [19]. SVR was included to account for potential nonlinear relationships between spectral features and SSC, consistent with recent fruit spectroscopy studies that apply support vector machine–based models alongside PLSR [2]. ETR was incorporated as an ensemble tree-based regression method. Ensemble learning approaches have been reported as increasingly applied in near-infrared spectroscopy combined with chemometrics methods [20].

Model hyperparameters were optimized using grid search combined with 10-fold cross-validation performed exclusively on the calibration subset. The same validation protocol, search ranges (Table I), and evaluation criteria were applied across preprocessing–model combinations and for both acquisition modes to maintain methodological consistency.

3) Model performance evaluation metrics: The predictive performance of regression models was evaluated using four commonly reported metrics: the correlation coefficient (R), coefficient of determination (R^2), root mean square error (RMSE), and residual prediction deviation (RPD). These metrics are widely adopted in NIR spectroscopy studies for fruit quality prediction and provide complementary perspectives on model accuracy [9], [10], [21].

The correlation coefficient R reflects the linear relation-

ship between reference and predicted SSC values, while the coefficient of determination R^2 indicates how much of the variability in the reference measurements is explained by the model. RMSE was calculated for both calibration ($\mathrm{RMSE_c}$) and prediction ($\mathrm{RMSE_p}$) sets as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

where, y_i and $\hat{y}_i$ denote reference and predicted SSC values, respectively, and n is the number of samples. RMSE provides a scale-dependent measure of prediction error and is commonly reported in fruit spectroscopy applications [2].

The residual prediction deviation RPD is frequently used in agricultural NIR modeling to assess practical predictive utility [3], [4]. It is calculated as the ratio between the standard deviation (SD) of the reference SSC values in the prediction set and the root mean square error of prediction ($\mathrm{RMSE_p}$) [22]:

$$\mathrm{RPD} = \frac{\mathrm{SD}}{\mathrm{RMSE_p}} \tag{6}$$

Higher RPD values indicate improved predictive performance. In NIR fruit quality studies, RPD is commonly used as an indicator of model applicability for screening or quantitative analysis.

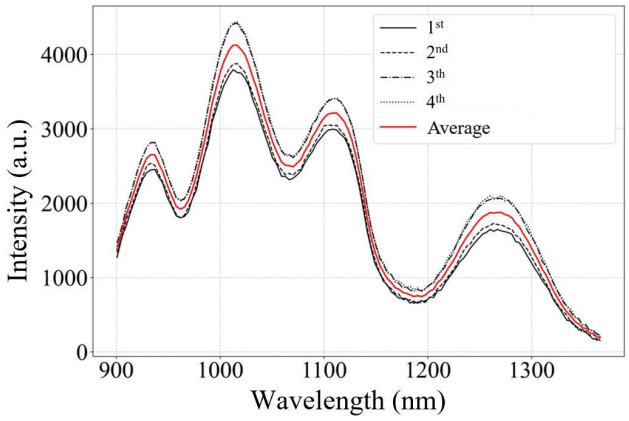
III. Results and Discussion

A. Spectral Consistency Under Consecutive Acquisitions

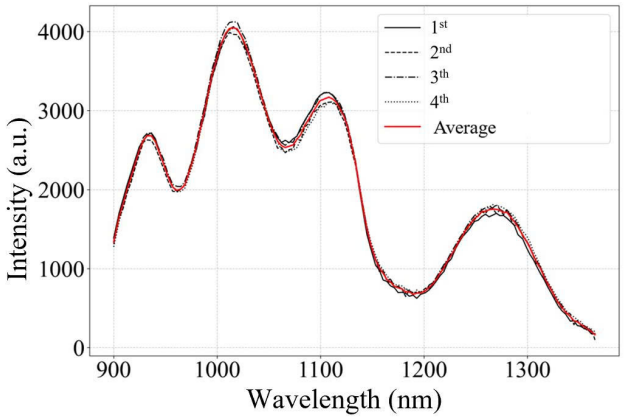
Fig. 5a presents four consecutive interactance spectra acquired at the same nominal measurement region using the manual method. Although the overall spectral profiles follow similar trends, substantial spectral dispersion among consecutive acquisitions is observed across multiple wavelength regions. Pronounced deviations are visible near the major peak around 1000 nm, as well as at secondary local maxima and minima between 1100 and 1300 nm. In several regions, the peak-to-peak separation among scans reaches several hundred a.u., indicating significant variability associated with manual probe repositioning and contact pressure. The calculated RMSD for the manual acquisition mode was approximately 1046 a.u., which represents the absolute spectral dispersion baseline for comparison, reflecting limited spectral consistency under manual operation.

In contrast, Fig. 5b shows four consecutive spectra obtained using the automated acquisition system. The spectra exhibit strong overlap throughout the measured wavelength range, including the local extrema between 1100 and 1300 nm. Only minor deviations are observed among acquisitions. The RMSD for the automated acquisition mode was approximately 363 a.u. Compared with the manual baseline of 1046 a.u., this corresponds to an absolute reduction of 683 a.u. and a relative reduction of approximately 65.3% in spectral dispersion.

These results indicate that the automated system provides improved spectral consistency under consecutive acquisitions, reflecting enhanced acquisition stability. The reduced spectral dispersion supports the reliability of the acquired data and is consistent with the regression performance observed under automated acquisition.



(a)



(b)

Fig. 5. Spectral consistency under consecutive acquisitions at the same nominal measurement region: (a) Manual acquisition and (b) Automated acquisition. Four consecutive scans are shown for each mode; the red curve represents the mean spectrum.

B. Regression Model Performance

Table II presents the calibration and prediction performance of PLS, SVR, and ETR models for mandarin SSC (°Brix) prediction under manual and automated interactance spectral acquisition modes using multiple preprocessing strategies. The objective of this analysis was not to identify an optimal SSC prediction model for grading applications, but rather to evaluate whether the automated interactance acquisition system provides spectral data suitable for chemometric modeling at a level comparable to manual measurement.

1) Effect of regression modeling approaches under different acquisition modes: Clear differences were observed between linear and nonlinear regression methods. Across preprocessing strategies, the PLS model achieved maximum prediction correlations of $R_\mathrm{p} = 0.68$ under manual acquisition and $R_\mathrm{p} = 0.74$

TABLE II. CALIBRATION AND PREDICTION PERFORMANCE OF PLS, SVR, AND ETR MODELS FOR MANDARIN SSC PREDICTION UNDER MANUAL AND AUTOMATED SPECTRAL ACQUISITION MODES USING DIFFERENT PREPROCESSING STRATEGIES.

Model	Acquisition mode	Preprocessing	Calibration			Prediction			
			R_c	R_c^2	$\mathrm{RMSE_c}$	R_p	R_p^2	$\mathrm{RMSE_p}$	RPD
PLS	Manual	MM	0.64	0.41	0.89	0.61	0.37	0.85	1.26
		SNV	0.67	0.45	0.88	0.68	0.38	0.66	1.35
		SG	0.60	0.36	0.92	0.67	0.42	0.84	1.33
		MMADN	0.64	0.41	0.91	0.59	0.27	0.76	1.23
		SG + FD	0.55	0.30	0.94	0.64	0.39	0.94	1.28
		SNV + SG	0.61	0.37	0.95	0.68	0.44	0.66	1.37
		SNV + SG + FD	0.59	0.35	0.94	0.63	0.37	0.79	1.29
		MMADN + SG + MM	0.55	0.30	0.96	0.65	0.36	0.89	1.29
		Unprocessed	0.65	0.42	0.88	0.56	0.26	0.91	1.20
	Automated	MM	0.57	0.33	0.98	0.59	0.32	0.90	1.24
		SNV	0.66	0.44	0.92	0.71	0.39	0.69	1.41
		SG	0.58	0.34	0.96	0.64	0.39	0.91	1.29
		MMADN	0.62	0.38	0.96	0.58	0.24	0.81	1.23
		SG + FD	0.52	0.27	0.99	0.65	0.38	0.96	1.28
		SNV + SG	0.60	0.36	0.98	0.67	0.37	0.71	1.35
		SNV + SG + FD	0.60	0.36	0.96	0.60	0.34	0.87	1.25
		MMADN + SG + MM	0.58	0.34	0.98	0.74	0.42	0.79	1.42
		Unprocessed	0.58	0.33	0.97	0.58	0.32	0.90	1.23
SVR	Manual	MM	0.82	0.67	0.66	0.68	0.46	0.79	1.36
		SNV	0.72	0.49	0.85	0.64	0.33	0.69	1.30
		SG	0.70	0.48	0.83	0.71	0.51	0.78	1.42
		MMADN	0.69	0.47	0.86	0.67	0.38	0.71	1.32
		SG + FD	0.71	0.49	0.81	0.73	0.52	0.83	1.44
		SNV + SG	0.66	0.44	0.89	0.72	0.50	0.62	1.43
		SNV + SG + FD	0.69	0.46	0.86	0.62	0.37	0.79	1.27
		MMADN + SG + MM	0.76	0.57	0.76	0.70	0.48	0.81	1.38
		Unprocessed	0.78	0.59	0.74	0.66	0.43	0.80	1.33
	Automated	MM	0.83	0.67	0.68	0.66	0.43	0.83	1.33
		SNV	0.71	0.49	0.88	0.69	0.44	0.67	1.38
		SG	0.69	0.47	0.86	0.76	0.58	0.76	1.54
		MMADN	0.71	0.48	0.88	0.68	0.43	0.70	1.35
		SG + FD	0.73	0.51	0.82	0.75	0.55	0.82	1.49
		SNV + SG	0.67	0.44	0.92	0.70	0.46	0.66	1.40
		SNV + SG + FD	0.70	0.48	0.87	0.58	0.33	0.87	1.22
		MMADN + SG + MM	0.77	0.58	0.78	0.78	0.56	0.69	1.57
		Unprocessed	0.79	0.60	0.76	0.65	0.41	0.84	1.31
ETR	Manual	MM	0.48	0.23	1.02	0.60	0.30	0.90	1.21
		SNV	0.99	0.96	0.25	0.67	0.27	0.72	1.34
		SG	0.94	0.85	0.45	0.85	0.66	0.65	1.74
		MMADN	0.98	0.93	0.32	0.61	0.27	0.77	1.20
		SG + FD	0.93	0.84	0.45	0.81	0.63	0.73	1.64
		SNV + SG	0.98	0.95	0.27	0.79	0.58	0.57	1.62
		SNV + SG + FD	0.98	0.94	0.29	0.65	0.38	0.78	1.31
		MMADN + SG + MM	0.92	0.82	0.49	0.77	0.55	0.75	1.50
		Unprocessed	0.63	0.38	0.91	0.66	0.34	0.86	1.27
	Automated	MM	0.47	0.22	1.05	0.55	0.25	0.95	1.18
		SNV	0.99	0.97	0.22	0.72	0.33	0.73	1.42
		SG	0.95	0.88	0.41	0.79	0.59	0.74	1.58
		MMADN	0.98	0.95	0.29	0.60	0.28	0.78	1.24
		SG + FD	0.95	0.87	0.42	0.81	0.63	0.74	1.67
		SNV + SG	0.94	0.83	0.50	0.78	0.52	0.62	1.56
		SNV + SG + FD	0.98	0.95	0.27	0.66	0.42	0.82	1.32
		MMADN + SG + MM	0.99	0.96	0.26	0.82	0.59	0.67	1.65
		Unprocessed	0.81	0.61	0.75	0.69	0.39	0.85	1.33

under automated acquisition. In comparison, SVR reached $R_{\mathrm{p}} = 0.73$ (manual) and $R_{\mathrm{p}} = 0.78$ (automated), while ETR achieved $R_{\mathrm{p}} = 0.85$ under manual acquisition and $R_{\mathrm{p}} = 0.82$ under automated acquisition.

These results indicate that nonlinear regression approaches generally provided higher prediction accuracy than the linear PLS model. However, the relative performance between manual and automated acquisition varied depending on the regression method. Automated acquisition yielded higher maximum R_{p} values for both PLS and SVR, whereas ETR exhibited a slight decrease in peak prediction performance compared with the manual mode.

Despite these differences, the overall ranking of regression methods remained consistent across acquisition modes. In both manual and automated measurements, nonlinear models outperformed PLS by a comparable margin. This consistency suggests that the automated acquisition configuration does not alter the fundamental modeling behavior observed with manually acquired spectra.

2) Influence of spectral preprocessing: Preprocessing strategy influenced prediction performance more strongly than acquisition mode. SG was associated with relatively higher R_{p} values in several configurations, particularly for nonlinear regression models. The highest prediction accuracy ($R_{\mathrm{p}} = 0.85$) was obtained using the ETR model with SG preprocessing under manual acquisition. Under automated acquisition, SG-based and combined preprocessing strategies also yielded strong performance, with R_{p} values reaching up to 0.82 for ETR and 0.78 for SVR when combined with additional scaling methods such as MMADN + SG + MM. These observations indicate that spectral smoothing contributes to improved signal stability and can enhance model performance, particularly for nonlinear regression methods.

SNV, commonly used for scatter correction, did not consistently produce higher prediction accuracy compared to smoothing-based approaches. Similarly, scaling procedures such as MM and MMADN affected model performance depending on the regression method and acquisition mode. No single preprocessing strategy uniformly outperformed others across all models, and the effectiveness of preprocessing remained dependent on the interaction between preprocessing method and regression algorithm.

Overall, the results suggest that preprocessing plays a more influential role in determining model performance than acquisition mode, with SG-based approaches providing consistently strong performance across both manual and automated measurements.

3) General synthesis and relevance to study objective: Across all regression approaches and preprocessing configurations, prediction performance under automated acquisition remained within the same range as that obtained under manual acquisition. Maximum R_{p} values and corresponding RPD values were comparable between acquisition modes for each regression method. In several preprocessing–model combinations, particularly PLS and SVR with SG-based preprocessing, slightly higher R_{p} values and lower $\mathrm{RMSE}_{\mathrm{p}}$ values were observed under automated acquisition. Because these differences were small and the objective of this study was to evaluate system applicability rather than superiority, no formal statistical significance analysis was performed. The results therefore support comparable predictive performance between the two acquisition modes.

Most RPD values fell between approximately 1.2 and 1.7, corresponding to exploratory or screening-level prediction capability, as RPD values below 2.0 are generally considered suitable only for rough screening or approximate predictions in NIR fruit quality modeling studies [4], [23]. Although this RPD range indicates limited predictive capability for grading applications [24], achieving grading-level accuracy was not the objective of this study. Overall, the results indicate that automated interactance spectral acquisition provides spectral data of comparable quality to manual measurement, while also offering improved acquisition stability and, in some configurations, slightly enhanced predictive consistency.

The consistent regression behavior and similar prediction ranges across acquisition modes indicate that the automated interactance system provides spectral data comparable to those obtained using manual measurements. Therefore, further improvements in prediction accuracy are more likely to depend on dataset size, sampling diversity, and calibration strategy rather than on acquisition configuration alone.

It should also be noted that this study employed conventional chemometric and machine learning models (PLS, SVR, and ETR) together with commonly used spectral preprocessing techniques to enable a transparent comparison between manual and automated acquisition modes. While these approaches are widely adopted in NIR spectroscopy studies, preprocessing-based modeling pipelines may have limitations in capturing complex nonlinear spectral relationships when larger datasets are available. Recent studies have shown that deep learning approaches, such as convolutional neural networks (CNNs), can automatically learn hierarchical spectral features and reduce reliance on handcrafted preprocessing [2]. With expanded datasets and more diverse sampling conditions, future research may therefore explore deep learning–based spectral modeling to further improve prediction accuracy while leveraging the stable spectral acquisition provided by the proposed automated system.

IV. Conclusion

This study presented and validated an automated interactance NIR spectral acquisition system for non-destructive prediction of mandarin SSC. Compared with the manual setup, the automated system reduced spectral variability among consecutive acquisitions within a measurement session by approximately 65%, confirming improved measurement stability under controlled mechanical positioning and repeatable probe–fruit contact conditions.

Regression models based on PLS, SVR, and ETR were developed using multiple preprocessing strategies under an identical cross-validation protocol to maintain methodological consistency between acquisition modes. Across all model configurations, prediction performance obtained from automated spectra remained within the same performance range as that derived from manual measurements, with several configurations showing slightly higher R_{p} values and lower $\mathrm{RMSE}_{\mathrm{p}}$ values under automated acquisition, particularly for PLS and SVR models combined with SG-based preprocessing.

The relative performance ranking of linear and nonlinear models was preserved across acquisition modes. These results indicate that improved acquisition stability did not adversely affect predictive performance and may be associated with small performance gains in certain configurations, while maintaining overall modeling consistency.

Although most RPD values corresponded to screening-level predictive capability, the primary objective of this study was feasibility validation rather than grading-level optimization. A direct comparison with commercial inline NIR systems in terms of cost, maintenance, and industrial scalability was beyond the scope of the present study; therefore, no quantitative claims are made regarding these aspects. Unlike previously reported online transmission-based systems, the proposed approach emphasizes interactance acquisition with controlled probe positioning and modular mechanical integration. The findings demonstrate that the proposed automated system provides stable spectral data suitable for chemometric modeling and supports future integration into automated fruit grading platforms.

While validated on mandarins, the mechanical positioning and optical acquisition architecture may be adapted to other fruits with similar structural properties. Further improvements in prediction accuracy are therefore more likely to depend on expanded datasets, improved sampling strategies, and advanced calibration methods rather than modifications to the acquisition mechanism alone.

ACKNOWLEDGMENT

This work was funded by the Vietnam Ministry of Education and Training under grant number B2024-TCT-20.

REFERENCES

[1] M. Mei and J. Li, "An overview on optical non-destructive detection of bruises in fruit: Technology, method, application, challenge and trend," *Computers and Electronics in Agriculture*, vol. 213, p. 108195, 2023.

[2] S. Zeng, Z. Zhang, X. Cheng, X. Cai, M. Cao, and W. Guo, "Prediction of soluble solids content using near-infrared spectra and optical properties of intact apple and pulp applying PLSR and CNN," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 304, p. 123402, 2024.

[3] Y. Ozaki, S. Morita, and Y. Morisawa, *Spectral Analysis in the NIR Spectroscopy*. Singapore: Springer Singapore, 2021, pp. 63–82.

[4] B. M. Nicolaï, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K. I. Theron, and J. Lammertyn, "Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review," *Postharvest Biology and Technology*, vol. 46, no. 2, pp. 99–118, nov 2007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925521407002293

[5] K. B. Walsh, J. Blasco, M. Zude-Sasse, and X. Sun, "Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use," *Postharvest Biology and Technology*, vol. 168, p. 111246, oct 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925521419303230

[6] Y. Yu, Q. Zhang, J. Huang, J. Zhu, and J. Liu, "Nondestructive determination of SSC in Korla fragrant pear using a portable near-infrared spectroscopy system," *Infrared Physics and Technology*, vol. 116, no. January, p. 103785, 2021. [Online]. Available: https://doi.org/10.1016/j.infrared.2021.103785

[7] Z. Hu, Y. Pu, W. Wu, L. Pan, Y. Yang, and J. Zhao, "Online detection of moldy apple core based on diameter and SSC features," *Food Control*, vol. 168, no. August 2024, p. 110879, feb 2025. [Online]. Available: https://doi.org/10.1016/j.foodcont.2024.110879 https://linkinghub.elsevier.com/retrieve/pii/S0956713524005966

[8] X. Tian, S. Fan, J. Li, W. Huang, and L. Chen, "An optimal zone combination model for on-line nondestructive prediction of soluble solids content of apple based on full-transmittance spectroscopy," *Biosystems Engineering*, vol. 197, pp. 64–75, sep 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1537511020301677

[9] Z. Wang, Q. Wang, J. Li, and W. Huang, "Non-destructive detection of soluble solids content in Shawo radish with spatial spectra extraction method based on the full transmission near-infrared spectroscopy," *Journal of Food Composition and Analysis*, vol. 148, no. P2, p. 108364, 2025. [Online]. Available: https://doi.org/10.1016/j.jfca.2025.108364

[10] C. Wang, X. Luo, Z. Guo, A. Wang, R. Zhou, and J. Cai, "Influence of the peel on online detecting soluble solids content of pomelo using Vis-NIR spectroscopy coupled with chemometric analysis," *Food Control*, vol. 167, no. July 2024, 2025.

[11] S. Kim, E. Lee, J. Park, J.-O. Nam, and S. R. Kim, "Evaluating the Nutritional Composition of Unripe Citrus and Its Effect on Inhibiting Adipogenesis and Adipocyte Differentiation," *Journal of Microbiology and Biotechnology*, vol. 34, no. 6, pp. 1206–1213, jun 2024.

[12] K. Timilsina and B. Tripathi, "Chemical quality attributes of mandarin (Citrus reticulata Blanco) as affected by altitude and fruit bearing position in Kavre, Nepal," *Archives of Agriculture and Environmental Science*, vol. 4, no. 3, pp. 319–325, 2019.

[13] R. K. H. Galvão, M. C. U. Araujo, G. E. José, M. J. C. Pontes, E. C. Silva, and T. C. B. Saldanha, "A method for calibration and validation subset partitioning," *Talanta*, vol. 67, no. 4, pp. 736–740, 2005.

[14] X. Wei, J. He, S. Zheng, and D. Ye, "Modeling for SSC and firmness detection of persimmon based on NIR hyperspectral imaging by sample partitioning and variables selection," *Infrared Physics and Technology*, vol. 105, no. October 2019, 2020.

[15] N. F. M. Kasim, P. Mishra, R. E. Schouten, E. J. Woltering, and M. P. Boer, "Assessing firmness in mango comparing broadband and miniature spectrophotometers," *Infrared Physics and Technology*, vol. 115, no. April, p. 103733, jun 2021.

[16] S. Gao and W. Xie, "SSC and pH prediction and maturity classification of grapes based on hyperspectral imaging," *Smart Agricultural Technology*, vol. 8, no. May, p. 100457, aug 2024. [Online]. Available: https://doi.org/10.1016/j.atech.2024.100457 https://linkinghub.elsevier.com/retrieve/pii/S2772375524000625

[17] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, dec 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1568494619302947

[18] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques," *TrAC Trends in Analytical Chemistry*, vol. 132, p. 116045, nov 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0165993620302740

[19] M. Fodor, A. Matkovits, E. L. Benes, and Z. Jókai, "The Role of Near-Infrared Spectroscopy in Food Quality Assurance: A Review of the Past Two Decades," *Foods*, vol. 13, no. 21, p. 3501, oct 2024. [Online]. Available: https://www.mdpi.com/2304-8158/13/21/3501

[20] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, apr 2006.

[21] S. Xu, H. Lu, Z. He, and X. Liang, "Non-destructive determination of internal soluble solid content in pomelo using visible/near infrared full-transmission spectroscopy," *Postharvest Biology and Technology*, vol. 214, p. 112990, aug 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925521424002357

[22] P.-L. Nguyen, T.-N. Pham, V.-G.-K. Tran, N.-T. Tran, M. Fukuzawa, C.-N. Nguyen, and C.-N. Nguyen, "Towards Automatic Mango Grading: A Computer Vision System Design for Measuring a Mango's Weight and Skin Defect Areas," *Vietnam Journal of Computer Science*, vol. 12, no. 02, pp. 153–170, may 2025. [Online]. Available: https://www.worldscientific.com/doi/10.1142/S2196888824400025

[23] C. M. Peraza-Alemán, A. López-Maestresalas, C. Jarén, J. I. Ruiz de Galarreta, L. Barandalla, and S. Arazuri, "Mapping acrylamide content in potato chips using near-infrared hyperspectral imaging and chemometrics," *Food Chemistry*, vol. 479, p. 143794, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308814625010453

[24] L. Kang, H. Luo, X. Ma, J. Yu, H. Liu, H. Liu, and Y. Tong, "Study on the effect of moisture content on the spectral detection of soluble solids in apricot," *Scientific Reports*, feb 2026. [Online]. Available: https://www.nature.com/articles/s41598-026-39890-w

The Impact of Modern AI on Software Development: A Systematic Literature Review

Meryem Bensaid[1]*, Marouane Achbari[2], Younesse Ouahbi[3], Soumia Ziti[4]
IPSS-Laboratory of Intelligent Processing and Security of Systems, Mohammed V University, Rabat, Morocco[1,2,3,4]
SMSD, Moroccan Society of Digital Health, Rabat, Morocco[4]
LIAS-Artificial Intelligence and Systems Laboratory, Hassan II University, Casablanca, Morocco[4]

***Abstract*—As large language models, and agentic AI systems are increasingly being integrated into software engineering, an expanding amount of empirical evidence surrounding these technologies has emerged. This systematic literature review examines the impact of modern AI techniques and tools in software development lifecycle phases and related activities, covering studies published between 2023 and 2025 and resulting in a corpus of 62 primary studies, investigating the role of large language models, AI agents and agentic AI workflows across lifecycle phases. The synthesis is guided by three research questions that address the entire software development cycle, reported impacts on development practices, outcomes, and constraints. This review fills a gap in the synthesis of modern AI applications and their impacts. It concludes that modern AI in software engineering is progressively evolving from a generation tool into a reasoning and coordination infrastructure layer, with ongoing efforts targeting the mitigation of identified limitations and the advancement of trustworthy agentic AI capabilities for dependable software engineering practices.**

***Keywords*—Large language models; generative AI; agentic AI; software development lifecycle; systematic literature review**

I. Introduction

Software development has changed significantly since large language models were released in late 2022. Targeted and customized automation is no longer the exclusive application of AI in this field. Every level of software development today benefits from the use of large language models, agents, and agent-based AI systems. These technologies now assist developers in defining requirements, making architectural decisions, writing code, fixing bugs, reviewing pull requests, detecting security risks, and managing production occurrences. As a result, AI is currently being used in majority of development workflows, tools, and infrastructures to help with software system design, implementation, testing, deployment, and maintenance.

Simultaneously, the quantity of empirical research on AI in software engineering has increased significantly. The majority of these studies concentrate on certain tools, jobs, or single lifecycle stages like program repair or code generation. They generate separate understandings even though they offer valuable insights. However, both researchers and practitioners still lack clarity on where modern AI works best, the risks it presents (especially with more autonomous agentic systems), and how its use should be governed. The lack of an integrated perspective makes it difficult to draw meaningful conclusions about the overall impact of modern AI on software development and the new organizational and technological challenges that come with it.

In order to address this gap, this study provides a comprehensive analysis of peer-reviewed primary research on the application of modern AI in software development published between 2023 and 2025 . The review adheres to the PRISMA 2020 guidelines and is organized around a 4-phase SDLC framework: Inception, Construction, Quality Assurance, and Operations. In total, 62 studies were selected and analyzed to address the three research questions. More specifically, the review examines how modern AI techniques are applied across SDLC activities, what impacts on software development are reported, and what challenges limit their effective adoption in practice. By bringing together evidence from different research streams, this study aims to provide a clear and structured cross-phase understanding of the evolving role of AI in software engineering.

The main contributions of this study are threefold. First, it provides a lifecycle-oriented synthesis of modern AI applications across software development activities. Second, it examines the stated impact of various technologies on productivity, artifact quality, and developer processes. Third, it outlines the primary limitations, threats, and adoption issues that affect the ethical integration of AI into software engineering practice.

The remainder of this study is organized as follows: Section II presents an overview of modern AI and the SDLC. Section III reviews related studies. Section IV discusses the research methodology. Section V provides results and addresses the research questions. Finally, Section VI concludes the study and outlines directions for future work.

II. Background

In this review, *modern AI* primarily refers to generative AI systems applied to software development, with a particular focus on large language model (LLM)-based tools, AI agents, and agentic workflows. Generative Artificial Intelligence (GenAI) denotes AI systems capable of producing original artifacts, including natural language text, images, and source code, in response to prompts or other forms of input [1]. In software engineering, the most relevant form of GenAI is represented by LLMs, which are deep learning models based on the Transformer architecture [2] and pre-trained on large-scale corpora of natural language and source code using self-supervised learning objectives. This paradigm enables broad

*Corresponding author

capabilities such as code generation, explanation, summarization, and reasoning support without requiring task-specific supervision [3].

Three levels of modern AI are distinguished throughout this review, where each one is characterized by an increasing degree of autonomy and coordination complexity: LLMs as reactive generation systems, AI Agents as tool-augmented planning architectures, and Agentic AI as multi-agent collaborative systems operating with limited human oversight. Table I summarizes these distinctions.

TABLE I. OPERATIONAL CHARACTERISTICS OF MODERN AI LEVELS

Level	Memory Type	Coordination	Role
LLM	Context-only	None	Text/code generation
AI Agent	Persistent	Single-agent	Tool-augmented task execution
Agentic AI	Shared	Multi-agent	Workflow orchestration

While LLMs provide strong generative and reasoning capabilities, a standard model usually operates as a reactive system that maps a given input to an output. To support more structured and autonomous task execution, recent research increasingly embeds LLMs into AI agents [4]. These agents consist of a base model combined with persistent memory, access to external tools, and planning or control mechanisms that enable multi-step perception–planning–action loops [5]. In software development, an AI agent can perform functions like interacting with software artifacts, running code execution tools, obtaining documentation, and iteratively refining outputs. Most recently, this paradigm has expanded toward *agentic AI*, in which one or more AI agents collaborate to achieve complex goals with limited human supervision [6]. In this case we find an orchestrating layer which typically divides a main goal into subtasks, distributes them to specialized agents, and consolidates intermediate results into a final output [7]. Compared to a single-call LLM interaction, agentic systems introduce higher degrees of autonomy, coordination, and decision delegation, making them well-suited for complex software engineering workflows. Fig. 1 illustrates this conceptual progression.

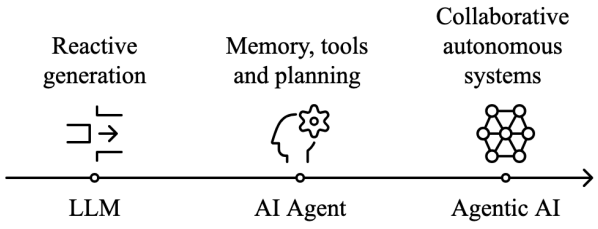


Fig. 1. Levels of modern AI in software development.

III. RELATED WORK

Recent reviews have examined modern AI in software engineering from several perspectives. Broad syntheses have mapped LLM applications across development tasks and highlighted challenges related to evaluation reliability, hallucination, and human–AI workflows [8], [9], while more focused studies have investigated specific subdomains such as software testing, vulnerability detection, and requirements engineering [10], [11], [12]. In parallel, dedicated reviews have explored multi-agent and agentic approaches, emphasizing autonomy, task orchestration, and collaboration patterns [7], [13].

However, these studies tend to focus on particular technologies or development activities, offering valuable but partial perspectives. A synthesis connecting how modern AI is applied across the full software development lifecycle, what impacts are reported, and what limitations and adoption challenges emerge from current evidence is still needed.

Accordingly, this study addresses that gap through a systematic literature review organized around a four-phase lifecycle framework: Inception, Construction, Quality Assurance, and Operations. Different from previous research such as [8] and [9], which discuss the use of LLMs without a lifecycle structure, and [7] and [13], which focus on agentic architectures without spanning all SDLC phases, this review synthesises empirical evidence on modern AI applications, reported impacts, and adoption challenges through a unified four-phase lifecycle lens.

IV. METHODOLOGY

The methodology section describes the systematic approach adopted in this study in order to ensure transparency, consistency, repeatability, and rigor in the review process. To achieve this, the review was conducted following the PRISMA guidelines [14] and aligns with the Kitchenham and Charters SLR guidelines for software engineering [15].

A. Study Objective and Scope

The aim of this systematic literature review is to synthesize empirical studies that examine how modern AI is being used in software development. The collected evidence is organized around three main aspects: first how AI is used across the software development lifecycle, what impacts on development are reported, and what limitations and adoption challenges are identified. The review focuses only on primary research studies, while secondary sources such as surveys, literature reviews, and mapping studies are excluded.

This review covers papers published between 2023 and 2025. This period was mainly chosen in order to include research conducted after the public release of ChatGPT which was late 2022 [16], and which created great interest in generative AI and large language models and their application in software development. By restricting the review to this time period, the study only focuses on actual advances in modern AI rather than classical or prior AI approaches.

To structure the synthesis across software development activities, this review adopts a consolidated four-phase analytical framework comprising *Inception*, *Construction*, *Quality Assurance*, and *Operations*. This structure is partially inspired by emerging industry conceptualizations of AI-driven development lifecycles, such as the AI-Driven Development Life Cycle (AI-DLC) [17]. However, unlike such models, this review treats Quality Assurance as a distinct analytical phase in order to better capture validation-oriented activities such as testing, debugging, code review, and security verification. Fig. 2 illustrates this framework and the main categories of activities associated with each phase.

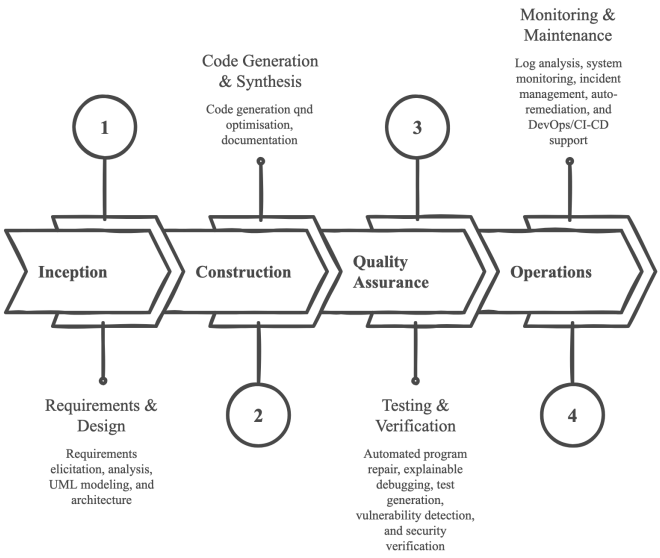


Fig. 2. Four-phase software development lifecycle framework adopted in the review.

TABLE II. PICOC FRAMEWORK DEFINING THE REVIEW SCOPE

Dimension	Definition
Population	Software developers, software engineering teams, researchers, and software projects.
Intervention	Modern AI approaches including generative AI, LLMs, AI code assistants, AI agents, retrieval-augmented generation, and agentic workflows.
Comparison	Non-AI approaches or alternative AI-based techniques, where applicable.
Outcome	SDLC applications, reported impacts, limitations, and adoption challenges.
Context	Peer-reviewed English-language studies in software engineering, published between 2023 and 2025.

To further define the scope and keywords used in our study we generated the PICOC framework presented in Table II.

B. Research Questions

To structure the synthesis of recent empirical evidence, this review is guided by three research questions addressing the usage, impact, and constraints of modern AI in software development.

RQ1: What are the current applications of modern AI in the SDLC?

This question aims to identify the current state and trends in the application of modern AI techniques throughout the software development lifecycle. By analyzing reported applications in activities such as requirements engineering, implementation, testing, and operational support.

RQ2: What are the reported impacts of modern AI on software development practices ?

This focuses on reported changes related to productivity, development speed, artifact quality, collaboration dynamics, and decision support.

RQ3: What limitations, risks and adoption challenges are perceived from the integration of modern AI in software development?

This research question addresses the technical reliability issues, security and governance concerns, organizational barriers, and the new challenges introduced by LLMs, multi-agent and autonomous AI systems.

C. Data Sources and Search Strategy

The literature search was mostly done on *Scopus*, which was used as the main data source due to its broad coverage of computer science topics. In order to achieve broader coverage of studies related to this research, *IEEE Xplore* and the *ACM Digital Library* were also used as they provide a vast amount of software engineering conference and journal papers. Preprint repositories such as *arXiv* were consulted to identify latest developments in generative AI, especially to inform the background and related work sections of this review. However, it should be noted that, only studies indexed in Scopus were retained in the final corpus, ensuring that all selected publications meet a peer-review quality standard [15].

The search string combined both modern AI and software development activity terms, as follows:

(*"generative AI" OR "large language model" OR "LLM" OR "AI agent" OR "agentic AI"*) AND (*"software development" OR "software engineering" OR "SDLC" OR "requirements engineering" OR "software architecture" OR "code generation" OR "software testing" OR "software maintenance" OR "DevOps"*)

The query was applied to the titles, abstracts, and keywords section. Additional filters were used to restrict results to the defined publication period (2023-2025), English language, document types corresponding to articles and conference papers, and the Computer Science subject area.

D. Inclusion and Exclusion Criteria

To ensure consistency in study selection, predefined inclusion and exclusion criteria were applied throughout the screening process. The inclusion criteria are summarized in Table III, and the exclusion criteria are presented in Table IV.

E. Study Selection Process

The study selection process followed a PRISMA-inspired workflow. The first phase of search involved using Scopus database and the predefined search string resulting in 5,336 records, which were then filtered to retain only English-language articles and conference papers published between 2023 and 2025 in the Computer Science subject area, reducing the corpus to 3,888 records.

These records were screened on the basis of titles and abstracts, retaining 399 studies for full-text assessment. Full-text screening was then conducted to verify compliance with the predefined inclusion and exclusion criteria and to confirm publication accessibility. After this final eligibility assessment, 62 studies were included in the review corpus, where 49 forming the main SDLC-oriented corpus used to address RQ1

TABLE III. INCLUSION CRITERIA BY SEARCH SUBJECT

Search Subject	Inclusion Criteria
Modern AI Applications in Software Development	– Proposes or evaluates LLM-based, generative AI, or agentic approaches applied to software development activities. – Addresses at least one SDLC phase: Inception, Construction, Quality Assurance, or Operations. – Reports empirical evaluation, controlled experiment, or tool validation.
Impacts on Development Practices and Outcomes	– Measures or discusses productivity, output quality, or workflow changes resulting from AI adoption. – Provides quantitative or qualitative evidence of AI impact on development speed, correctness, or developer support. – Highlights strengths, limitations, and practical implications of the proposed approach.
Limitations, Risks, and Adoption Challenges	– Identifies technical, security, or organizational barriers to AI adoption in software engineering. – Discusses reliability issues such as hallucination, non-determinism, or adversarial robustness. – Focuses on governance, privacy, or socio-technical challenges in AI-assisted development. – Includes specific findings on agentic AI risks such as cascading errors or coordination failures.
Publication Eligibility	– Published between 2023 and 2025. – Peer-reviewed primary journal article or conference paper. – Written in English and retrieved within the defined database search scope.

TABLE IV. EXCLUSION CRITERIA BY SEARCH SUBJECT

Search Subject	Exclusion Criteria
Scope and Domain	– Not focused on software development activities (e.g. healthcare, education, or finance). – Modern AI is not the primary intervention or technique under study. – Uses only traditional ML, rule-based systems, or classical NLP without LLMs or generative AI.
Empirical Evidence	– Lacks sufficient empirical evidence or without experimental validation. – Secondary study such as a survey, systematic literature review, systematic mapping, or meta-analysis.
Publication Accessibility	– Full text not accessible for eligibility assessment or data extraction.

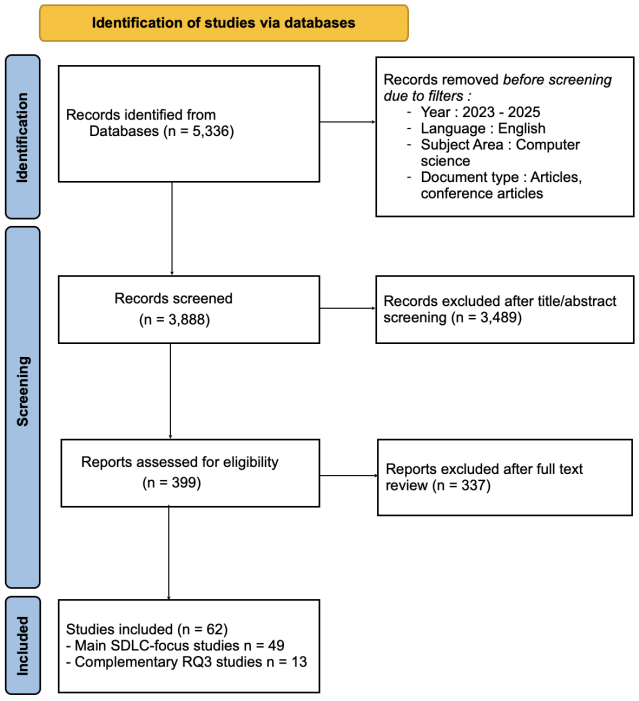


Fig. 3. PRISMA flow diagram for the review.

and RQ2, and 13 complementary studies included to strengthen the analysis of limitations, risks, and adoption challenges in RQ3. The overall selection process is summarized in Fig. 3.

V. RESULTS AND DISCUSSION

A. RQ1: Applications of Modern AI Across SDLC Phases

AI tools are currently employed throughout all stages of software development, from requirements gathering to system maintenance, as shown in Table V.

1) Inception: In Inception phase, AI is mostly used for requirements engineering, software modeling, and architectural assistance. LLMs help in generating requirements from conversations held with clients [21], refine backlog items and improve user story quality [19], [25], or detect ambiguities or inconsistencies in specifications [23], [29]. They can also help generate UML diagrams [22], [24], [26], and help novice analysts work with modeling tools [28]. Overall, AI in Inception mostly assists with drafting early outputs and formalizing requirements, whereas planning and estimation remain underexplored.

2) Construction: In Construction phase, AI is used for a wide range of tasks including writing code and generating documentation. Rather than asking a model to generate code from a simple bare prompt, studies show that methods like retrieval-augmented generation (RAG), guided prompting, or requirement-aware pipelines significantly improve the correctness of the generated code [34], [36], [38]. When developers need to handle more complex tasks, iterative refinement and multi-agent systems help improve results further [35], [45], [46]. Recent work also shows that AI can generate code directly from visual inputs such as flowcharts and UML diagrams [40], [41], and that LLMs can be embedded into CI/CD pipelines to automatically convert backlog items into functional code [39]. AI additionally supports documentation

TABLE V. SYNTHESIS OF REVIEWED STUDIES ACROSS SOFTWARE DEVELOPMENT PHASES.

Ref.	Phase	Sub-theme	AI Approach
[18]	Inception	Requirements elicitation	GPT-4-Turbo agent simulation
[19]	Inception	Requirements elicitation	Multi-agent LLM system
[20]	Inception	Backlog classification	RoBERTa fine-tuning
[21]	Inception	Requirements from conversations	Llama-2 prompting
[22]	Inception	UML and code generation	LLM-driven MDA pipeline
[23]	Inception	SRS quality evaluation	GPT-4 zero-shot
[24]	Inception	UML sequence diagrams	GPT-3.5 prompting
[25]	Inception	User story quality	LLM agent prompting
[26]	Inception	UML class diagram modeling	GPT-4 evaluation
[27]	Inception	Domain model generation	ChatGPT-3.5
[28]	Inception	UML modeling assistance	LLM support for novices
[29]	Inception	Requirements ambiguity detection	GPT-4 zero-shot
[30]	Inception	Requirements elicitation	ChatGPT prompting
[31]	Inception	Interview script generation	ChatGPT and Bard
[32]	Inception	Domain model extraction	GPT-3.5 vs CRF comparison
[33]	Inception	Architectural decision records	GPT-4 and Flan-T5
[34]	Construction	RAG-based code generation	Retrieval-augmented generation
[35]	Construction	Iterative code refinement	Multi-role LLM pipeline
[36]	Construction	Guided code generation	BM25 retrieval prompting
[37]	Construction	Decoding configuration	Temperature tuning
[38]	Construction	Requirement-to-code generation	Requirement-aware prompting
[39]	Construction	Backlog-to-code pipeline	GPT-4 CI/CD integration
[40]	Construction	Flowchart-to-code generation	Multimodal LLMs
[41]	Construction	UML-to-code generation	GPT-4 Vision
[42]	Construction	REST API endpoint extraction	LLM agents
[43]	Construction	Use case documentation	Structured prompting
[44]	Construction	Code summarization	LLM-based summarization
[45]	Construction	End-to-end agentic development	Multi-agent framework (MetaGPT)
[46]	Construction	End-to-end agentic development	Chat chain protocol (ChatDev)
[47]	Quality Assurance	Automated program repair	LLM & static analysis
[48]	Quality Assurance	Automated program repair	Retrieval-augmented repair
[49]	Quality Assurance	Multi-method bug repair	Divide-and-conquer agents
[50]	Quality Assurance	Explainable debugging	Scientific debugging prompting
[51]	Quality Assurance	Vulnerability-witnessing tests	GPT-4 Turbo
[52]	Quality Assurance	Test quality assessment	GPT-4o and Amazon Q
[53]	Quality Assurance	Automated test generation	State space models
[54]	Quality Assurance	Smart contract vulnerability	GPT-3.5 fine-tuned
[55]	Quality Assurance	Smart contract vulnerability	Multi-agent audit system
[56]	Quality Assurance	Industrial code review	Fine-tuned LLM (BitsAI-CR)
[57]	Quality Assurance	Code review automation	Prompt engineering vs fine-tuning
[58]	Operations	Cloud observability	Knowledge graph & LLM
[59]	Operations	Network monitoring	Llama-3.2 multi-agent
[60]	Operations	Incident root cause analysis	Hierarchical multi-agent (SRE)
[61]	Operations	Maintenance assistance	Ontology-based RAG
[62]	Operations	Incident mitigation	Fine-tuned Code-davinci
[63]	Operations	Automated root cause analysis	GPT-4 workflow-driven (RCACopilot)
[64]	Operations	Log parsing	Flan-T5 and LLaMA few-shot
[65]	Operations	Auto-remediation	GPT-4 Ansible playbook generation
[66]	Operations	CI/CD pipeline support	Workflow-integrated LLM

tasks such as generating use case descriptions and summarizing source code [43], [44], while LLM-based agents have been used to extract REST API endpoints automatically from microservice architectures [42].

3) Quality assurance: In Quality Assurance phase, AI is mainly used for code review, automated testing, vulnerability recognition, debugging, and software repair. Pairing LLMs with static analysis [47], retrieval [48], or multi-agent decomposition [49] results in more reliable bug fixes compared to direct generation. AI debugging tools assist developers in determining whether a fix is correct [50]. AI may generate syntactically valid test cases [51], [52], but without class-level dependencies or runtime data, the tests frequently lack semantic correctness [53]. In order to improve vulnerability detection fine-tuned [54] and multi-agent [55] approaches can be used in domains such as smart contracts, but generalization to larger codebases remains limited. In code review, the use of prompt engineering techniques [56] and fine-tuning [57] provide more meaningful and reliable feedback than basic prompting.

4) Operations: In Operations phase, AI can assist in doing log analysis, giving incident diagnosis and maintenance guidance. Systems based on LLM can semantically interpret logs [64]. Now, incident management has progressed from single-model diagnostic support [62] to multi-agent [59], [60] and workflow-driven systems [63] that integrate logs, metrics, and traces to identify root causes and suggest solutions. In addition, AI can also generate remediation scripts and playbooks through maintenance assistants [61] and auto-remediation frameworks [65]. More recently, LLMs can also help developers work with GitHub repositories by reading configuration files and supporting CI/CD decisions [66]. Overall, AI in Operations is acting as an intelligent interpretation and action layer, though scalability, safety, and performance in complex production environments remain open challenges.

Across phases, the role of AI in software engineering has gradually increased, moving from helping structure early outputs in Inception, to generating and coordinating code in the Construction stage, supporting verification in Quality Assurance, and taking action in Operations. However, fully autonomous AI remains currently out of reach due to limitations in semantic reasoning, domain knowledge, and integration complexity.

Looking across the reviewed studies, several patterns emerge beyond what Table V individually captures. While contributions are present across all lifecycle phases, their distribution within each phase is uneven. In Inception, studies are heavily concentrated on requirements elicitation and UML modeling, leaving activities such as project planning and effort estimation largely unaddressed. Construction and Quality Assurance attract the broadest range of contributions, as tasks like code generation, program repair, and testing are naturally suited to automation. Operations, despite growing interest, remains the least explored phase, particularly regarding runtime safety and long-term system maintenance.

In addition to these distributions, certain patterns in the effectiveness of modern AI methodologies have been observed. For example, the use of retrieval-augmented generation (RAG) generally increases the accuracy of output compared to zero-shot or basic prompting methods, especially in tasks related to code generation and repair due to the fact that it grounds model outputs in project-specific context and reduces the chances of hallucination. Furthermore, the use of multi-agent and agentic systems demonstrate better performance in complex, multi-step workflows. However, such systems need more coordination and suffer from a higher risk of cascading mistakes. Finally, fine-tuning provides smoother and specialized performance especially when it comes to code reviewing and vulnerability detection but involves high resource expenses. Overall, none of the discussed methods outperforms others in all cases, and the efficiency of techniques depends greatly on the complexity and nature of the task.

B. RQ2: Reported Impacts on Software Development Practices and Outcomes

The reviewed studies report three main categories of impact associated with the adoption of modern AI: faster development, more consistent outputs, and changes in how development teams work.

The most commonly reported benefit of modern AI adoption in software development is increasing productivity, as it helps reduce manual work in requirements elicitation [18], [19] and backlog management [25]. AI also produces more correct code especially when used with retrieval [34], guided prompting [36], or generation guided by project requirements [38]. In Quality Assurance and Operations, AI speeds up program repair [47], code review [56], incident diagnosis [60], log parsing [64], and root cause recommendations [62]. That said, these gains are task-dependent and tend to drop in complex or open-ended activities [23], [40], [53].

Multi-agent systems also changed how development teams work, as they support task decomposition [45], [46] and connect AI to CI/CD and operational workflows [65], [66]. LLM tools can also help junior developers with modeling [28], debugging [50], and incident response [61].

As for output quality, AI tools provide better results particularly in requirements [24], documentation [43], [44], and code review [56]. However, a recurring issue that has been found in the literature is that these outputs are often syntactically correct but semantically wrong, meaning that the code compiles and follows proper syntax rules but does not actually do what was intended or fails to capture the required behavior, especially in testing [51] and vulnerability detection [54], [55]. This is an issue that persists across all lifecycle phases and is addressed by introducing several techniques. Fig. 4 summarizes the three main approaches repeatedly reported in the reviewed studies: prompt engineering, retrieval-augmented generation, and fine-tuning. Prompt engineering structures the model's reasoning before it generates, leading to more accurate code [36], better debugging support [50], and more useful code review feedback [56], then we have Retrieval-augmented generation that grounds the model in relevant existing code or documentation, with improvements reported in code generation [34], program repair [48], and maintenance assistance [61]. Finally, fine-tuning, which is a technique that adjusts the model according to the task or domain, enhancing its effectiveness for software repair [49], automated code review [57], and vulnerability detection [54]. Together, these three techniques consistently outperform basic zero-shot generation by highlighting the fact that raw model capability alone is insufficient for AI-assisted development.

Taken together, the literature suggests that modern AI contributes to incremental yet systemic shifts in software development practices. While productivity acceleration and artifact standardization are increasingly observable, fully autonomous development remains constrained by semantic reasoning limitations, domain specificity, and integration challenges.

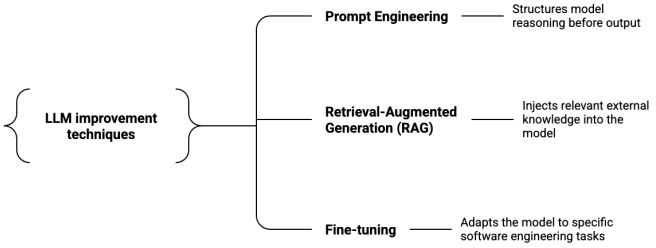


Fig. 4. LLM-based techniques for improving output quality in AI-assisted software development.

C. RQ3: Limitations, Risks, and Adoption Challenges

A broad set of limitations and risks has been encountered in the literature, as summarized in Table VI. These challenges span technical reliability issues, architectural constraints, and socio-technical adoption barriers, and are further amplified in emerging agentic development paradigms.

TABLE VI. COMPLEMENTARY RQ3 STUDIES

Ref.	Focus	Risk Category
[67]	GenAI adoption in organizations	Governance
[68]	Uncertainty in code completions	Automation bias
[69]	LLM self-verification	Self-contradiction
[70]	Correctness via internal states	Reasoning limits
[71]	Data race detection	Runtime visibility
[72]	Security of AI-generated code	Vulnerabilities
[73]	LLM security awareness	Passive security risk
[74]	Membership inference attacks	Privacy leakage
[75]	Adversarial code completion	Robustness
[76]	Multi-agent process modeling	Coordination errors
[77]	Debate-verify-debug pipelines	Cascading errors
[78]	LLM agent security paradigms	Chained attacks
[79]	End-to-end multi-agent generation	Autonomy risks

From a technical perspective, hallucination and non-determinism are the most consistently reported limitations across SDLC phases. As discussed in RQ2, AI-generated outputs are frequently syntactically plausible yet semantically incorrect, affecting requirements modeling [23], code generation [35], program repair [49], and operational diagnosis [58]. Variability in outputs for identical inputs complicates reproducibility and quality assurance, while decoding configuration [37] and prompt sensitivity [44] further increase deployment complexity. Adversarial robustness also remains limited: recent studies demonstrate that malicious natural-language instructions can significantly increase the likelihood of insecure code generation while preserving functional correctness [75]. In agentic settings, cascading errors across multi-step pipelines introduce additional instability, as early planning inaccuracies propagate through downstream development stages [38], [77], [79].

Architectural and resource constraints represent a second category of limitations. Context window restrictions hinder cross-file reasoning in requirements analysis [29], program repair [47], and vulnerability detection [54], leading most studies to evaluate AI performance on simplified task units. Multi-agent systems also come with a lot of costs: they consume more tokens, introduce coordination delays, and require careful

orchestration, both in agentic development frameworks [45] and operational environments [60]. Approaches such as fine-tuning [57], retrieval integration [34], and hybrid verification pipelines [53] further demand significant computational resources and labelled data, which limits how widely they can be adopted. Another gap is that most current AI tools work on static code or logs and cannot observe system behavior at runtime, leaving dynamic issues [71] and vulnerabilities that only appear during execution [55] largely undetected.

Security and organizational challenges add another layer of concern. AI-generated code frequently contains vulnerabilities [72], and models rarely flag security issues on their own unless the user explicitly asks [73]. There is also a growing risk around privacy where attackers can use membership inference techniques to extract information about the data used to train a model, potentially exposing proprietary code or sensitive information [74]. In agentic systems, the risks go further — chained attacks that combine prompt injection with other exploitation techniques can manipulate how agents use tools or communicate with each other, creating new threats that traditional security models were not designed to handle [78].

Aside from these technological issues, organizational challenges also arise. Accepting AI suggestions without verification can become a harmful habit for developers [68]. Also, there is no clear legal framework governing how AI tools handle sensitive code or data [67]. While AI can assist junior engineers learn and speed up repetitive processes, over-reliance on low-quality or imprecise outputs [56], [61] might silently decrease the standard of the final result.

VI. Conclusion and Future Work

In conclusion, this review shows that current software development tasks tend to incorporate modern artificial intelligence. In accordance with the reviewed literature, not only does AI play the role of providing individual help, but it tends to assist in development processes overall. It is considered as an infrastructure layer that facilitates developers' work on various software development stages. Its advantages are primarily perceived in automated, repetitive and structured actions. However, there are still considerable challenges to AI integration, such as security problems, privacy protection, context management limitations, and other factors.

For practitioners, this review highlights various insights that can actually be taken into consideration. In this regard, prompt engineering, RAG, and fine-tuning consistently beat zero-shot approaches in terms of performance and need to be considered an integral part of the development process and not just additional features. As for multi-agent systems, they provide performance boosts for more intricate activities, but come with coordination costs and error propagation, which need proper management.

Future studies should focus on various gaps found in the current literature, most specifically in project planning, CI/CD pipeline automation, repository-level development, and vulnerability detection. The effects of AI adoption on developer skills, team structures, and engineering culture also presents a relevant investigation. As multi-AI agents and agentic AI systems become more and more common, we should also consider how agent teams should be organized and scaled as task complexity increases, on whether AI can support genuine creativity and innovation beyond pattern reproduction, and how autonomous agents can make reliable decisions when outputs conflict or ambiguities arise. Finally, as agentic AI becomes more capable, advancements in abilities must be accompanied by equally significant improvements in safety, accountability, and governance.

References

[1] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative AI," *Business & Information Systems Engineering*, vol. 66, pp. 111–126, 2024.

[2] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] C. Ebert and P. Louridas, "Generative AI for Software Practitioners," *IEEE Software*, vol. 40, no. 4, pp. 30–38, 2023.

[4] L. Wang et al., "A Survey on Large Language Model Based Autonomous Agents," *Frontiers of Computer Science*, vol. 18, no. 6, 2024.

[5] Y. Liu et al., "Agent Design Pattern Catalogue," *Journal of Systems and Software*, 2025.

[6] IBM, "What is Agentic AI?," 2024.

[7] J. He, C. Treude, and D. Lo, "LLM-Based Multi-Agent Systems for Software Engineering," *ACM Transactions on Software Engineering and Methodology*, 2025.

[8] X. Hou et al., "Large Language Models for Software Engineering: A Systematic Literature Review," *ACM Transactions on Software Engineering and Methodology*, 2024.

[9] A. Fan et al., "Large Language Models for Software Engineering: Survey and Open Problems," in *Proc. ICSE-FoSE*, 2023.

[10] J. Wang et al., "Software Testing with Large Language Models: Survey, Landscape, and Vision," *IEEE Transactions on Software Engineering*, 2024.

[11] X. Zhou et al., "Large Language Model for Vulnerability Detection and Repair: Literature Review and the Road Ahead," *ACM Transactions on Software Engineering and Methodology*, 2025.

[12] A. Hemmat et al., "Research Directions for Using LLM in Software Requirement Engineering: A Systematic Review," *Frontiers in Computer Science*, 2025.

[13] N. Akbar et al., "Methods and Techniques of Agentic Software Engineering: A Systematic Literature Review," *IEEE Access*, 2026.

[14] M. J. Page et al., "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews," *BMJ*, vol. 372, 2021.

[15] B. Kitchenham and S. Charters, "Guidelines for Performing Systematic Literature Reviews in Software Engineering," EBSE Technical Report EBSE-2007-01, 2007.

[16] OpenAI, "ChatGPT: Optimizing Language Models for Dialogue," OpenAI Blog, Nov. 2022.

[17] S. P. Raja and A. Mishra, "AI-Driven Development Life Cycle (AI-DLC)," AWS DevOps Blog, 2025.

[18] Ataei et al., "Elicitron: An LLM Agent-Based Simulation Framework for Design Requirements Elicitation," 2024.

[19] Sami et al., "Early Results of an AI Multiagent System for Requirements Elicitation and Analysis," 2024.

[20] Van Can and Dalpiaz, "Locating Requirements in Backlog Items: Content Analysis and Experiments with Large Language Models," 2025.

[21] Voria et al., "RECOVER: Toward Requirements Generation From Stakeholders' Conversations," 2025.

[22] Babaalla et al., "LLM-Driven MDA Pipeline for Generating UML Class Diagrams and Code," 2025.

[23] Krishna et al., "Using LLMs in Software Requirements Specifications: An Empirical Evaluation," 2024.

[24] Ferrari et al., "Model Generation with LLMs: From Requirements to UML Sequence Diagrams," 2024.

[25] Zhang et al., "LLM-based Agents for Automating the Enhancement of User Story Quality: An Early Report," 2024.

[26] De Bari et al., "Evaluating Large Language Models in Exercises of UML Class Diagram Modeling," 2024.

[27] Bozyigit et al., "Generating Domain Models from Natural Language Text Using NLP: A Benchmark Dataset and Experimental Comparison of Tools," 2024.

[28] Wang et al., "How LLMs Aid in UML Modeling: An Exploratory Study with Novice Analysts," 2024.

[29] Mahbub et al., "Can GPT-4 Aid in Detecting Ambiguities, Inconsistencies, and Incompleteness in Requirements Analysis? A Comprehensive Case Study," 2024.

[30] Ronanki et al., "Investigating ChatGPT's Potential to Assist in Requirements Elicitation Processes," 2023.

[31] Görer and Aydemir, "Generating Requirements Elicitation Interview Scripts with Large Language Models," 2023.

[32] Arulmohan et al., "Extracting Domain Models from Textual Requirements in the Era of Large Language Models," 2023.

[33] Dhar et al., "Can LLMs Generate Architectural Design Decisions? An Exploratory Empirical Study," in *IEEE International Conference on Software Architecture (ICSA)*, 2024.

[34] Yang et al., "An Empirical Study of Retrieval-Augmented Code Generation: Challenges and Opportunities," *ACM Transactions on Software Engineering and Methodology*, 2025.

[35] Chang et al., "A Self-Iteration Code Generation Method Based on Large Language Models," 2024.

[36] Li et al., "AceCoder: An Effective Prompting Technique Specialized in Code Generation," *ACM Transactions on Software Engineering and Methodology*, 2024.

[37] Arora et al., "Optimizing LLMs for Code Generation: Which Hyperparameter Settings Yield the Best Results?," 2024.

[38] Han et al., "ARCHCODE: Incorporating Software Requirements in Code Generation with Large Language Models," 2024.

[39] Sarschar et al., "PACGBI: A Pipeline for Automated Code Generation from Backlog Items," in *ASE*, 2024.

[40] He et al., "Flow2Code: Evaluating Large Language Models for Flowchart-based Code Generation Capability," 2024.

[41] Antal et al., "Toward a New Era of Rapid Development: Assessing GPT-4-Vision's Capabilities in UML-Based Code Generation," in *LLM4Code*, 2024.

[42] Chaplia and Klym, "Extracting REST API Endpoints from Microservices Using LLM Agents," in *DESSERT*, 2024.

[43] Naimi et al., "Automating Software Documentation: Employing LLMs for Precise Use Case Description," in *KES*, 2024.

[44] Sun et al., "Source Code Summarization in the Era of Large Language Models," 2024.

[45] Hong et al., "MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework," in *ICLR*, 2024.

[46] Qian et al., "ChatDev: Communicative Agents for Software Development," in *ACL*, 2024.

[47] Li et al., "Hybrid Automated Program Repair by Combining Large Language Models and Program Analysis," 2024.

[48] Liu et al., "ReAPR: Automatic Program Repair via Retrieval-Augmented Large Language Models," *Software Quality Journal*, 2025.

[49] Xie et al., "PReMM: LLM-Based Program Repair for Multi-method Bugs via Divide and Conquer," in *OOPSLA*, 2025.

[50] Kang et al., "Explainable Automated Debugging via Large Language Model-driven Scientific Debugging," 2023.

[51] Antal et al., "Leveraging GPT-4 for Vulnerability-Witnessing Unit Test Generation," in *EASE*, 2025.

[52] Alves et al., "Quality Assessment of Python Tests Generated by Large Language Models," 2024.

[53] Iznaga et al., "Integrating Large Language Models into Automated Software Testing," *Future Internet*, 2025.

[54] Liu et al., "Exploring the Potential of ChatGPT in Detecting Logical Vulnerabilities in Smart Contracts," *Blockchain: Research and Applications*, 2025.

[55] Wei et al., "Advanced Smart Contract Vulnerability Detection via LLM-Powered Multi-Agent Systems," *IEEE Transactions on Software Engineering*, 2025.

[56] Sun et al., "BitsAI-CR: Automated Code Review via LLM in Practice," 2024.

[57] Pornprasit and Tantithamthavorn, "Fine-tuning and Prompt Engineering for LLMs-based Code Review Automation," *Information and Software Technology*, 2024.

[58] Amanmadov and Abdullayev, "From Logs to Knowledge: LLM-Powered Dynamic Knowledge Graphs for Real-Time Cloud Observability," 2025.

[59] Yoon et al., "OFCnetLLM: Large Language Model for Network Monitoring and Alertness," 2025.

[60] Kataria, "Intelligent Site Reliability Engineering: A Multi-agent LLM Framework for Automated Incident Analysis and Root Cause Determination," *IJES*, 2025.

[61] Ludwig et al., "An Ontology-based Retrieval Augmented Generation Procedure for a Voice-Controlled Maintenance Assistant," *Computers in Industry*, 2025.

[62] Ahmed et al., "Recommending Root-Cause and Mitigation Steps for Cloud Incidents using Large Language Models," in *ICSE*, 2023.

[63] Chen et al., "Automatic Root Cause Analysis via Large Language Models for Cloud Incidents (RCACopilot)," in *EuroSys*, 2024.

[64] Ma et al., "LLMParser: An Exploratory Study on Using Large Language Models for Log Parsing," in *ICSE*, 2024.

[65] Sarda et al., "Leveraging Large Language Models for the Auto-remediation of Microservice Applications," in *FSE Industry Track*, 2024.

[66] Zhang et al., "On the Effectiveness of LLMs for GitHub Workflows," 2024.

[67] Kemell et al., "Adoption of Generative AI in Software Organizations: A Multiple Case Study," 2025.

[68] Vasconcelos et al., "Uncertainty Highlighting for AI Code Completions," 2024.

[69] Yu et al., "Fight Fire With Fire: Self-Verification Mechanisms in LLM-Based Code Generation," 2024.

[70] Bui et al., "Correctness Assessment of AI-Generated Code via Internal Representations," 2024.

[71] Alsofyani and Wang, "Evaluating ChatGPT for Data Race Detection," 2024.

[72] Tihanyi et al., "How Secure is AI-Generated Code?," 2024.

[73] Sajadi et al., "Do Large Language Models Consider Security? An Empirical Study," 2024.

[74] Yang et al., "GOTCHA: Membership Leakage Attacks Against Code Models," 2023.

[75] Wan et al., "Adversarial Attacks on Code Completion Systems via Natural Language Injection," 2024.

[76] Lin et al., "MAO: Multi-Agent Orchestration for Process Modeling," 2025.

[77] Islam et al., "DVDCoder: Debate-Verify-Debug Multi-Agent Framework for Code Generation," 2024.

[78] Gasmi et al., "Security Analysis of LLM Agent Systems: MCP vs Function Calling," 2025.

[79] Mao et al., "Blueprint2Code: A Multi-Agent Framework for End-to-End Software Development," 2025.

A Survey on Extended Reality Technologies for Industrial Maintenance

Mouad Danane, Abdelmajid El Ouadi, Youssef Rochdi
National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco

***Abstract*—Industrial maintenance increasingly relies on extended reality (XR) technologies, yet comprehensive analysis comparing AR, VR, and MR implementations with human-centric evaluation remains limited. This systematic literature review employs the PRISMA methodology to analyze 95 primary studies (2014–2025) that investigate implementation patterns, benefits, challenges, and evaluation criteria across XR modalities. AR dominates operational guidance (71.6%), VR prevails in training (38.9%), and MR enables collaborative maintenance (6.3%). Temporal analysis reveals three evolutionary phases: basic visualization (2014–2017), cognitive enhancement (2018–2021), and AI-integrated adaptive systems (2022–2025), mirroring the transition from Maintenance 4.0 to 5.0. Demonstrated benefits include 38% task completion time reductions and 92.4% error decreases. Human-centric factors appear in 44.2% of studies overall, with temporal analysis showing a progressive increase from approximately 25% in 2014–2017 to 58% in 2022–2025, substantiating a paradigm shift toward prioritizing cognitive support and user experience. Two novel frameworks advance theoretical understanding: an Evaluation Framework that expresses effectiveness as a function of technology, task complexity, and human factors; and a Technology-Task-Context Alignment Model that prescribes optimal XR-maintenance pairings. Critical gaps include the need for longitudinal field studies, standardized evaluation protocols, and economic analyses.**

***Keywords*—Industrial maintenance; maintenance 5.0; extended reality; augmented reality; virtual reality; mixed reality; human-centric design; artificial intelligence; systematic literature review**

I. Introduction

Manufacturing industries worldwide face mounting pressure to optimize operational efficiency while managing increasingly complex equipment and production systems [88]. Industrial maintenance, encompassing inspection, repair, diagnosis, and preventive activities, is essential for maintaining uninterrupted production, reliable equipment performance, and safe working conditions [89]. The advent of Industry 4.0, marked by the convergence of digital-physical systems, IoT technologies, AI, and sophisticated data analytics, has fundamentally transformed maintenance paradigms [90]. Despite these technological advances, human operators remain essential in overseeing, supervising, and servicing automated infrastructure, tackling issues beyond the scope of automation capabilities [91]. This industrial evolution seeks to augment human operator performance and minimize human error through intelligent, human-centered systems [92].

Extended Reality (XR)—which includes Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR)—has become a powerful enabler for industrial maintenance [86]. These immersive technologies offer distinct capabilities: AR overlays digital information onto physical environments, enabling real-time guidance during maintenance operations; VR creates fully immersive virtual environments for risk-free training and simulation; MR blends digital and physical elements to facilitate collaborative problem-solving [93]. The convergence of XR with advanced industrial technologies, including IoT sensors, digital twins, and AI-driven analytics, creates unprecedented opportunities to improve maintenance efficiency, training effectiveness, and knowledge transfer [94]. The demonstrated benefits include substantial reductions in task completion time, significant decreases in failure rates, enhanced mental support for complex procedures, and improved knowledge retention during training [95].

Despite substantial research attention, critical gaps persist in our understanding of XR application in industrial maintenance. Existing reviews have primarily focused on individual XR modalities, without adopting comprehensive approaches that compare AR, VR, and MR. In addition, limited attention has been paid to human-centric design principles and systematic evaluation methodologies, which are increasingly recognized as critical to successful implementation. The rapid technological evolution—particularly post-2020 advances in hardware capabilities, AI integration, and software platforms—necessitates updated analysis reflecting current state-of-the-art systems. Finally, the emerging Maintenance 5.0 framework, which emphasizes human-AI collaboration and sustainable practices, requires examination of how XR technologies facilitate this paradigm shift.

Several systematic reviews have previously examined aspects of this domain. Palmarini et al. [83] conducted an early review focused exclusively on AR in maintenance, identifying fragmented solutions and hardware limitations. Guo et al. [84] reviewed VR applications across product lifecycles. More recently, Wang et al. [82] examined XR in product assembly, and Alam et al. [86] analyzed XR challenges in Industry 4.0 contexts. However, none of these reviews comprehensively compare all three XR modalities (AR, VR, and MR) while simultaneously addressing human-centric design factors, evaluation methodologies, and the emerging Maintenance 5.0 framework. This gap motivates the present study.

This systematic review fills these research voids through an in-depth examination of XR technologies in industrial maintenance through examination of 95 primary studies published between 2014 and 2025. The review adopts a holistic approach that encompasses AR, VR, and MR technologies, while emphasizing human-centric design factors aligned with the Maintenance 5.0 principles. This work offers multiple novel contributions that advance the current literature. First, it provides updated empirical evidence of XR adoption patterns,

technological trends, and implementation outcomes reflecting recent advances. Second, it systematically analyzes evaluation methodologies employed across studies, identifying relationships between technology characteristics, task complexity, and appropriate assessment criteria. Third, it examines human-centric enhancement strategies, documenting how XR technologies address cognitive load, ergonomics, adaptability, and user experience. Fourth, it develops theoretical frameworks that advance understanding of XR effectiveness in maintenance contexts.

To achieve a comprehensive analysis, this review addresses five research questions: XR implementation patterns, benefits and potential, challenges and prospects, human-centric enhancements, and evaluation criteria for assessing these technologies in industrial maintenance contexts. These questions guide a systematic examination of the current state, demonstrated outcomes, implementation barriers, and future directions for XR in maintenance applications.

The subsequent sections of this study are organized in the following manner. Section II outlines the systematic review methodology based on PRISMA guidelines. Section III presents results organized by technology prevalence, application areas, advantages, limitations, human-centric factors, and evaluation criteria. Section IV discusses the findings in the context of theoretical frameworks, contrasts the outcomes with prior studies, and describes the practical applications and study limitations. Section V presents concluding remarks, principal findings, and prospective research avenues.

II. Methodology

A. Research Questions

The aim of this systematic literature review (SLR) is to provide a comprehensive overview of the application of Extended Reality (XR) technologies in industrial maintenance, including the primary tasks, key advantages, limitations, and challenges. Additionally, this study seeks to identify the evaluation criteria used to assess XR technologies in industrial maintenance and explore how XR can promote human-centricity, a key concept in the Maintenance 5.0 framework. Furthermore, this SLR aims to provide researchers with insights into future work and research directions. To achieve these objectives, the following research questions (RQs) have been formulated:

- RQ1: What is the current state of XR technologies in industrial maintenance?
- RQ2: What are the benefits and potential of XR technologies in industrial maintenance?
- RQ3: What challenges and prospects exist for XR technologies in industrial maintenance?
- RQ4: How can XR technologies enhance human-centric industrial maintenance?
- RQ5: What are the evaluation criteria used in the literature to assess XR technologies in industrial maintenance?

B. Research String and Databases Selection

To answer the research questions, a comprehensive search string was formulated to capture relevant studies on the application of XR technologies in industrial maintenance. The search string used was: "industrial maintenance" AND ("virtual reality" OR "augmented reality" OR "mixed reality" OR "extended reality"). This search string covers the main concepts of the research topic, including industrial maintenance and the various types of XR technologies (virtual reality, augmented reality, mixed reality). The literature search was conducted using three major scientific databases: Scopus, ScienceDirect, and Web of Science. These databases were selected for their relevance and extensive coverage of peer-reviewed literature in the fields of engineering, technology, computer science and energy.

To guarantee that the review captured current and pertinent research on the topic, the scope was restricted to publications from January 1, 2014, and May 19, 2025. Additionally, only studies published in English were considered to maintain consistency and accessibility for the researchers. These date range and language filters were applied directly in the database search phase, before proceeding to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol.

The initial search yielded a total of 321 articles, distributed as follows:

- ScienceDirect: 168 articles.
- Scopus: 86 articles.
- Web of Science: 67 articles.

C. Selection of Primary Study

The initial search yielded 321 articles from the selected databases. These articles were then subjected to the PRISMA protocol (Fig. 1) to identify the most relevant studies for the systematic literature review.

First, duplicate articles were removed, resulting in 255 unique articles. These articles were then screened using their titles, abstracts, and keywords to assess their relevance to the research questions. Articles that were out of scope like "6G networks: Beyond Shannon towards semantic and goal-oriented communications" and "Deep Learning-Powered System for Real-Time Digital Meter Reading on Edge Devices," were excluded. However, articles addressing extended reality in industrial maintenance, including assembly and disassembly tasks, were retained, as these tasks are considered integral to industrial maintenance. This screening process resulted in the exclusion of 124 articles, leaving 131 articles for further review.

The remaining 131 articles are reviewed in full text and subjected to detailed analysis to assess their eligibility for inclusion in our study. During this stage, an additional 36 articles were excluded for the following reasons:

- Upon closer examination, 10 articles were found to be not directly relevant to the research questions.

- 14 articles were identified as literature review articles, which did not meet the inclusion criteria for primary studies.
- 12 articles were not accessible for full-text review.

Finally, a total of 95 articles were selected as primary studies to answer the research questions. These articles form the basis for the data extraction and synthesis stages of the systematic literature review.

D. Quality Assurance and Limitations of the Search Strategy

Several methodological considerations should be noted regarding the PRISMA implementation. First, the search was limited to three academic databases (Scopus, ScienceDirect, Web of Science), and grey literature sources such as technical reports, white papers, and industry documentation were not included. This may introduce publication bias, as studies reporting positive outcomes are more likely to be published in peer-reviewed venues. Second, the study screening and selection process was conducted by the first author and verified by the co-authors through discussion-based consensus rather than formal inter-rater reliability assessment. While this approach is common in systematic reviews within engineering disciplines, it represents a methodological limitation. Third, no formal risk-of-bias assessment tool (e.g., ROBINS-I or the Newcastle-Ottawa Scale) was applied to individual primary studies, as the heterogeneity of study designs—ranging from controlled experiments to case studies and framework proposals—made standardized quality scoring impractical. These limitations are acknowledged and should be considered when interpreting the synthesized findings.

E. Co-Authorship Analysis

Co-authorship analysis was performed using VOSviewer on the 95 primary study papers selected for the systematic literature review on extended reality (XR) in industrial maintenance. The analysis aimed to identify influential authors and collaborative networks within this specific subset of literature.

From the total pool of authors in the 95 papers, a minimum threshold of two documents per author was set. Out of the authors who met this threshold, the top 50 were selected for visualization. The resulting co-authorship network (see Fig. 2) reveals several distinct clusters of collaborating authors. The most prominent cluster includes Hakulinen, J., Burova, A., Ronkainen, K., Keskinen, T., and Mäkelä, J., forming a densely connected red network at the center of the visualization. Other notable clusters involve the blue grouping of Fiorentino, Michele and Gattullo, Michele in the upper right, and the green cluster with Bai, Xiaoliang, Liu, Wei, and He, Weiping in the upper center. Additional collaborations are visible with Al-harkan, I.M. and Alhag, M.H. (yellow nodes), as well as isolated researchers like Romero, I. and Yu, H. distributed around the periphery.

The size of each author's node represents their relative influence and contribution to the primary studies, based on the number of co-authored publications within this subset of literature. The links between authors signify collaborative relationships. The proximity of authors within the visualization suggests thematic or institutional similarities in their research within the scope of this review.

F. Keywords Co-occurrence Analysis

A co-occurrence analysis of keywords from the 95 primary studies was conducted using VOSviewer to identify the main themes in XR research for industrial maintenance. With a minimum threshold of three occurrences, 25 out of 564 keywords were selected for visualization.

The co-occurrence network (see Fig. 3) reveals clusters centered around "augmented reality," "industrial maintenance," and related concepts such as "maintenance operations," "human computer interaction," and "industry 4.0." The strong connection between "augmented reality" and "industrial maintenance" highlights the significance of AR in this domain. By using VOSviewer, the keyword "augmented reality" is the most frequent keyword (56 occurrences), followed by " maintenance" (34), " industrial maintenance" (33), and "virtual reality" (29). These results underscore the pivotal importance of AR and VR technologies in industrial maintenance research.

The analysis provides an overview of key themes and their interconnections, underlining the importance of human factors, performance optimization, and the integration of XR with Industry 4.0. This offers critical perspectives for understanding prevailing research patterns and identifying promising areas for investigation in XR for industrial maintenance.

III. Extended Reality in Industrial Maintenance: Research Results

A. Literature Analysis

1) Publications per year: Fig. 4 shows the number of selected primary papers on XR in industrial maintenance from 2014 to 2025. There is an overall upward trend, indicating growing research interest. The count of publications increased from 1 paper in 2014 to a peak of 13 papers in 2018. This was followed by a slight decline to 10 papers in both 2019 and 2020. The year 2021 saw a resurgence with 12 papers, followed by a decrease to 8 papers in 2022. Research output rebounded to 10 papers in 2023 and peaked again at 13 papers in 2024, matching the previous high from 2018. The data for 2025 shows only two papers, as the data collection period ended early in the year. The trend suggests sustained interest in XR research for industrial maintenance over the past decade, with consistently significant output from 2018 onwards despite some year-to-year fluctuations.

2) Publications per database: Fig. 5 illustrates the distribution of the selected primary papers across three databases: Scopus, ScienceDirect, and Web of Science. ScienceDirect contributes the largest portion of the papers at 51.6% (49 papers) of the total collection. Scopus accounts for a similar but slightly smaller proportion at 46.3% of the papers (44 papers). Web of Science appears to be minimally represented, accounting for only 2.1% of the total (2 papers). However, this underrepresentation can be attributed to the duplication removal process conducted using Zotero. A considerable number of papers from Web of Science were merged with their duplicates from Scopus and ScienceDirect. The decision to prioritize the Scopus and ScienceDirect versions of duplicated

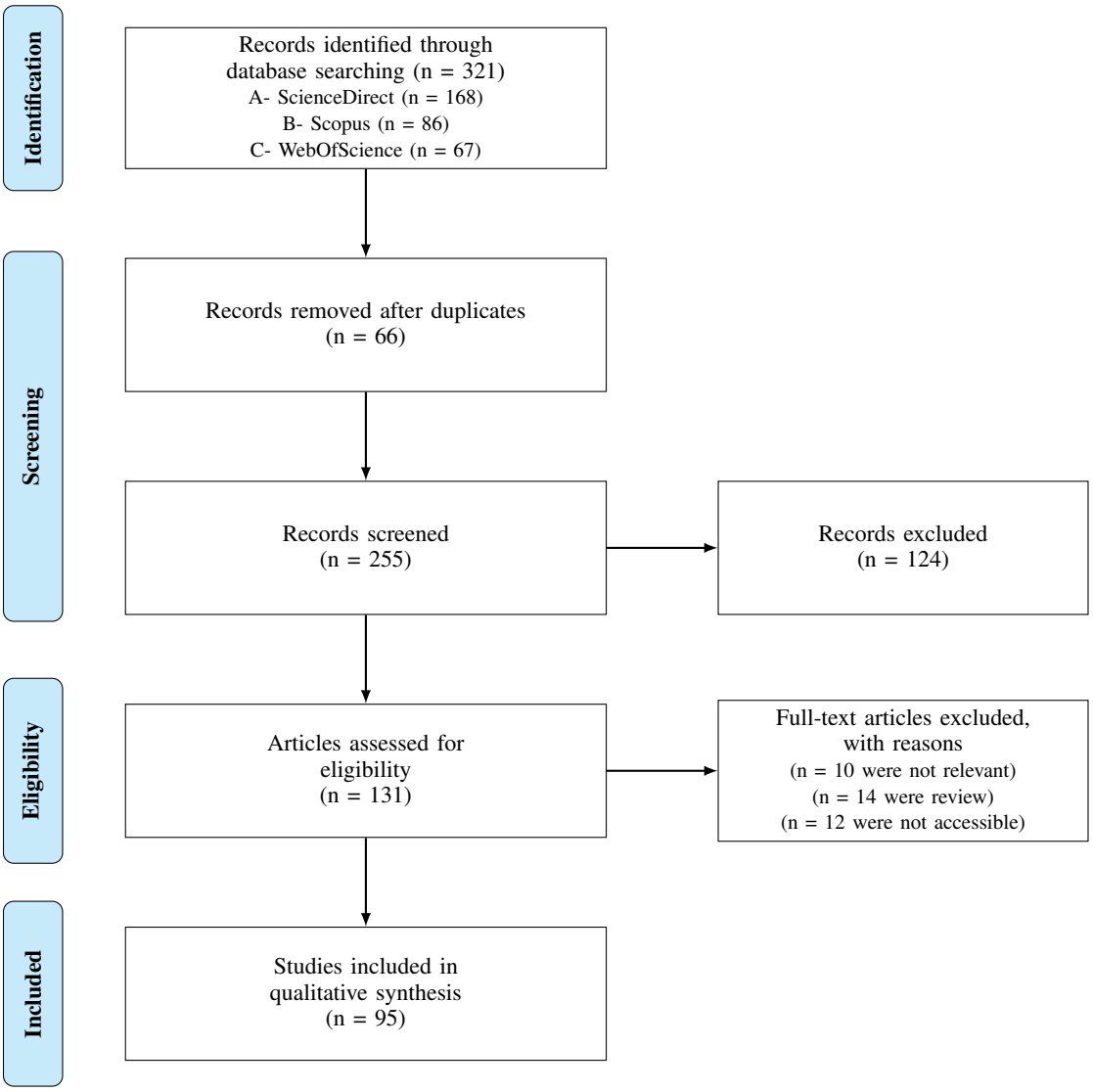


Fig. 1. The selection and filtering of primary studies (PRISMA flow diagram).

papers was based on the fact that these databases provide more extensive metadata, which facilitates and accelerates the analysis process. Consequently, the low percentage of papers from Web of Science does not necessarily reflect the database's actual contribution to the research on XR in industrial maintenance.

3) Publications per reference type: Fig. 6 illustrates the distribution of primary papers by reference type, categorized into journal articles, conference papers, and book chapters. Journal articles constitute the clear majority at 73.7% of the total, reflecting a strong preference for publishing XR research in peer-reviewed journals, which are generally seen as having a more rigorous review process and greater impact. Conference papers account for 25.3% of the publications, providing a platform for the rapid dissemination of knowledge and scholarly interaction. Book sections make up a minor portion at just 1.1% of the total. This distribution highlights the diverse avenues through which XR research in industrial maintenance is disseminated within the scientific community, while underscoring the predominance of journal publications in this field.

4) Publications per publication venues: Fig. 7 shows the distribution of papers across the top 10 publication venues. The most prominent sources include:

- Applied Sciences (Switzerland) journal (6 papers)
- Computers in Industry journal (6 papers)
- Advanced Engineering Informatics journal (4 papers)
- Lecture Notes in Computer Science Series (4 papers)
- Robotics and Computer-Integrated Manufacturing journal (4 papers)
- 16th IFAC Symposium on Information Control Problems in Manufacturing (3 papers)

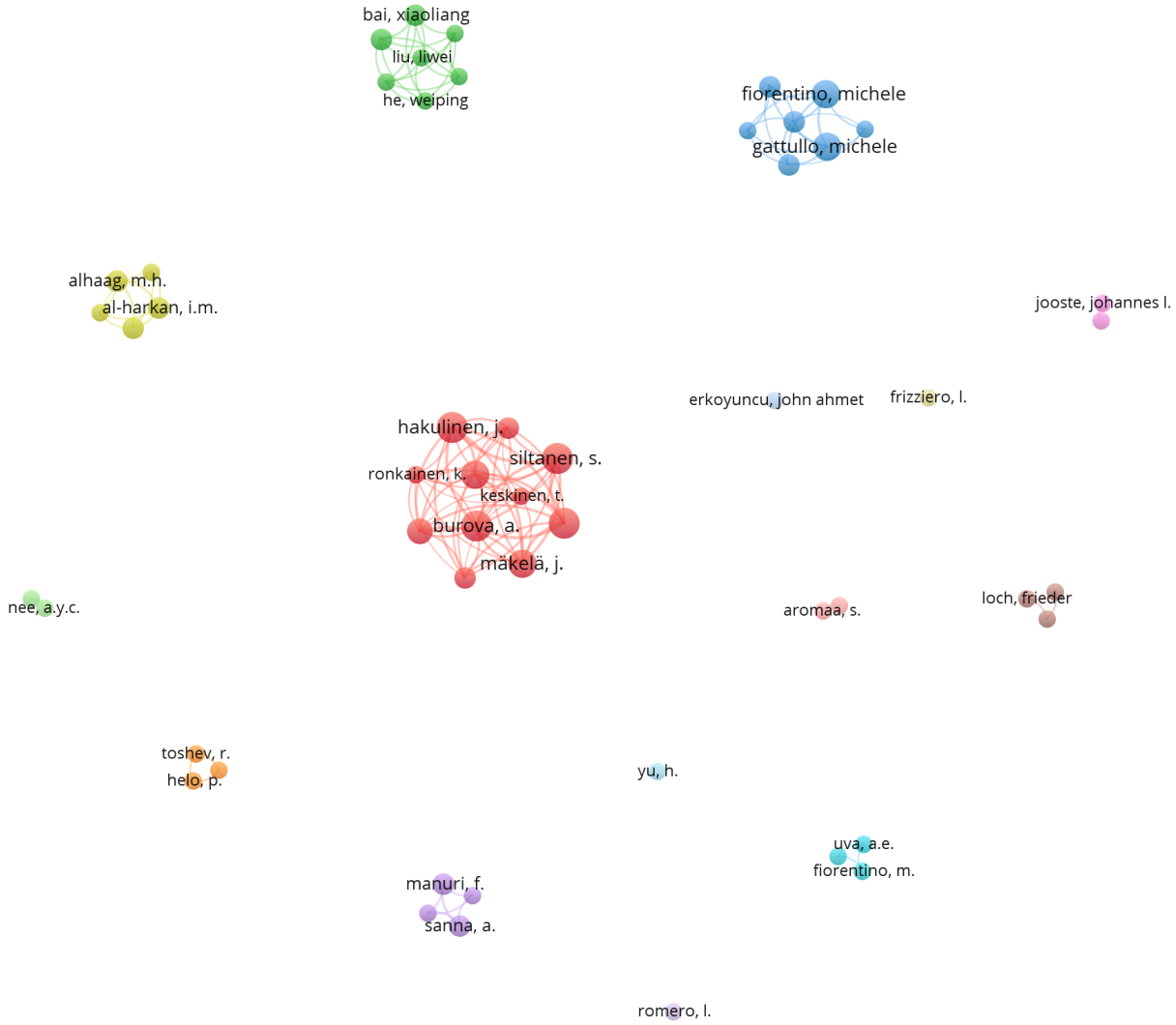


Fig. 2. Bibliometric map of co-authorship from VOSviewer using author names.

- 27th International Conference on Production Research (3 papers)
- 56th CIRP International Conference on Manufacturing Systems (2 papers)
- ACM International Conference Proceeding Series (2 papers)
- Applied Ergonomics journal (2 papers)

The distribution reveals that both specialized industrial and manufacturing journals, as well as computer science publications, serve as essential platforms for XR in industrial maintenance research. The relatively even distribution across these top venues indicates a multidisciplinary research landscape, with papers appearing in publications covering diverse areas, including applied sciences, computer science, engineering informatics, robotics, manufacturing, and ergonomics.

B. Extended Reality in Industrial Maintenance

1) Current state of extended reality in industrial maintenance: The analysis of 95 studies reveals a diverse and rapidly evolving landscape of extended reality (XR) technologies in industrial maintenance. This subsection outlines the key findings regarding the current state of these technologies.

a) Prevalence and distribution of XR technologies: The analysis of XR technology adoption across the 95 primary studies reveals clear patterns in the prevalence and distribution of different reality technologies in industrial maintenance applications. Fig. 8 presents the overall distribution of XR technologies identified in the literature.

Augmented Reality (AR) emerges as the dominant technology, appearing in 68 studies (71.6% of the total), making it the most widely adopted XR technology for industrial maintenance applications. Studies [2], [3] showcase AR's effectiveness in providing visual instructions for maintenance tasks, while studies [4], [5] highlight its use in enhancing worker performance and reducing errors. Virtual Reality (VR) is the second-most

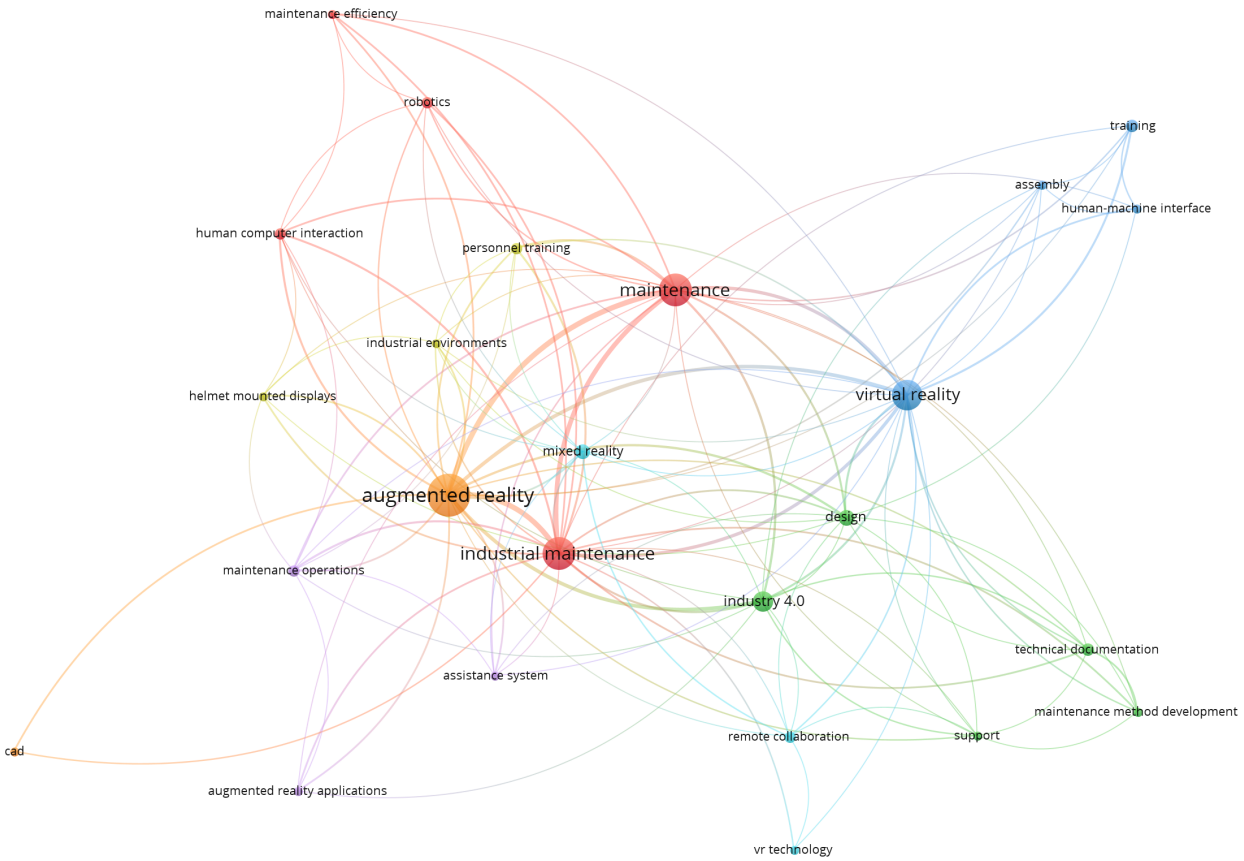


Fig. 3. Bibliometric map of author keywords from VOSviewer.

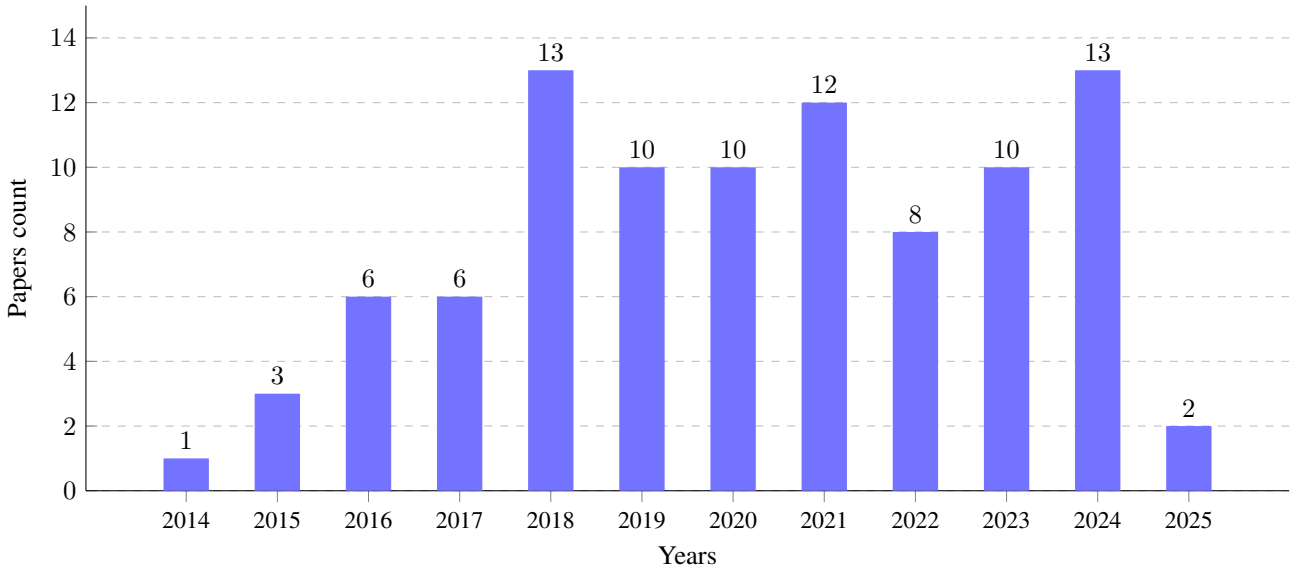


Fig. 4. Number of primary papers per year.

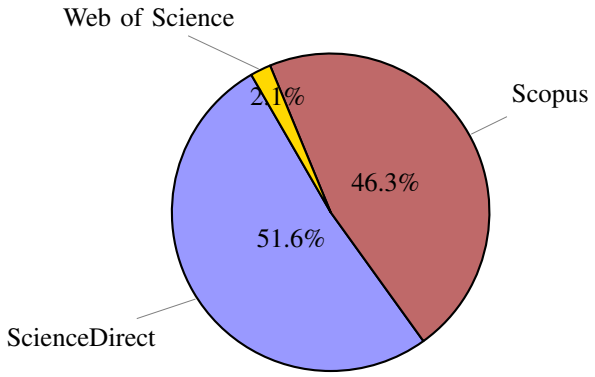


Fig. 5. Number of primary papers per database.

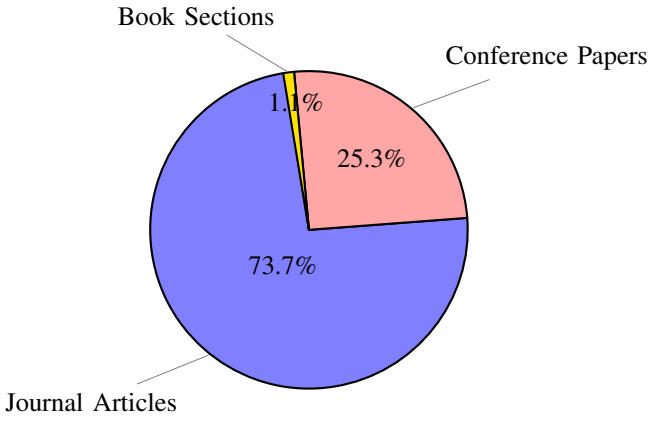


Fig. 6. Number of primary papers per reference type.

prevalent technology, used in 37 studies (38.9%). Studies [10], [12], [13] illustrate VR's capacity to create immersive, risk-free environments for maintenance personnel to practice complex procedures. Mixed Reality (MR) appears in only 6 studies (6.3%). Studies [68], [70], [71] demonstrate MR's potential to bridge the gap between on-site technicians and remote experts, facilitating real-time guidance and problem-solving.

A small number of papers (3, representing 3.1%) explicitly use the term "Extended Reality" (XR) in their work, typically in studies that present conceptual frameworks or methodologies applicable across multiple reality technologies. 1 paper in our dataset mentioned augmented virtuality as technology.

It is worth noting that 14 studies (14.7%) employed multiple XR technologies within the same research, with the most common combination being AR+VR (7 studies). Notably, three studies [10], [34], [70] incorporated all three main technologies (AR+VR+MR) in their work, with study [34] integrate them along with explicit XR framework development. This trend toward multi-technology approaches highlights the complementary nature of different XR modalities and suggests a growing interest in comprehensive XR ecosystems for maintenance applications.

C. Temporal Trends in XR Technology Adoption

The chronological analysis of XR technology adoption reveals significant trends in the field's evolution (Fig. 9). The three evolutionary phases were delineated based on a combination of qualitative thematic shifts and observable inflection points in the data. The boundary between Phase 1 and Phase 2 (2017–2018) corresponds to the sharp increase in publication volume (from 6 to 13 papers) and the emergence of cognitive enhancement as a research focus, coinciding with the commercial availability of Microsoft HoloLens and its rapid adoption in industrial AR studies. The boundary between Phase 2 and Phase 3 (2021–2022) aligns with the appearance of AI-integrated systems and automatic content generation approaches in the literature, reflecting the broader convergence of XR with machine learning and IoT technologies. These boundaries are therefore informed by convergent technological, thematic, and bibliometric indicators rather than purely statistical criteria. Early research (2014-2017) showed a strong preference for AR, which dominated the initial studies in industrial maintenance. Study [22] represents the first comprehensive AR implementation for maintenance tasks in our dataset, focusing on providing visual guidance on a screen-based video see-through AR system for motorcycle engine maintenance.

From 2017 onward, there was a noticeable increase in VR adoption, with powerful representation in 2018, 2021, and 2024. The years 2021-2022 mark a potential inflection point, where VR studies temporarily equaled or exceeded AR studies in quantity. Study [10] exemplifies this trend, presenting a VR-based methodology for assessing cognitive conditions during assembly training.

Mixed Reality, representing the most recent technological evolution in the reality continuum, first appeared in our dataset in 2020 [8] and has shown consistent but limited adoption through 2024. This slower uptake may reflect the nascent state of MR technology and its development ecosystem compared to more established AR and VR platforms.

The most recent years (2023-2025) have seen a renewed interest in AR applications, a diversification of technology approaches with more MR studies, and an increasing mention of XR as an overarching concept. This evolution suggests the field is moving toward a more integrated view of reality technologies while maintaining a strong interest in application-specific implementations.

D. Technology Distribution by Research Focus

The prevalence of specific XR technologies correlates strongly with the primary research focus and application context. AR dominates in studies focused on real-time guidance [15], [21], [27], [41], [43], maintenance instructions [22], [25], [26], [36], and on-site assistance [5], [7], [9]. This aligns with AR's fundamental capability to overlay digital information on real-world equipment, making it particularly suited for in-situ maintenance support.

VR is predominantly employed in studies focused on training and simulation of maintenance procedures [10], [30], [35], [48], [50], [63], [67], [64] and collaborative design reviews [11], [13], [32]. The immersive nature of VR creates controlled, risk-free environments that are ideal for skill development and scenario-based learning before interacting with physical equipment.

Mixed Reality applications appear most frequently in studies focusing on remote collaboration and tasks requiring so-

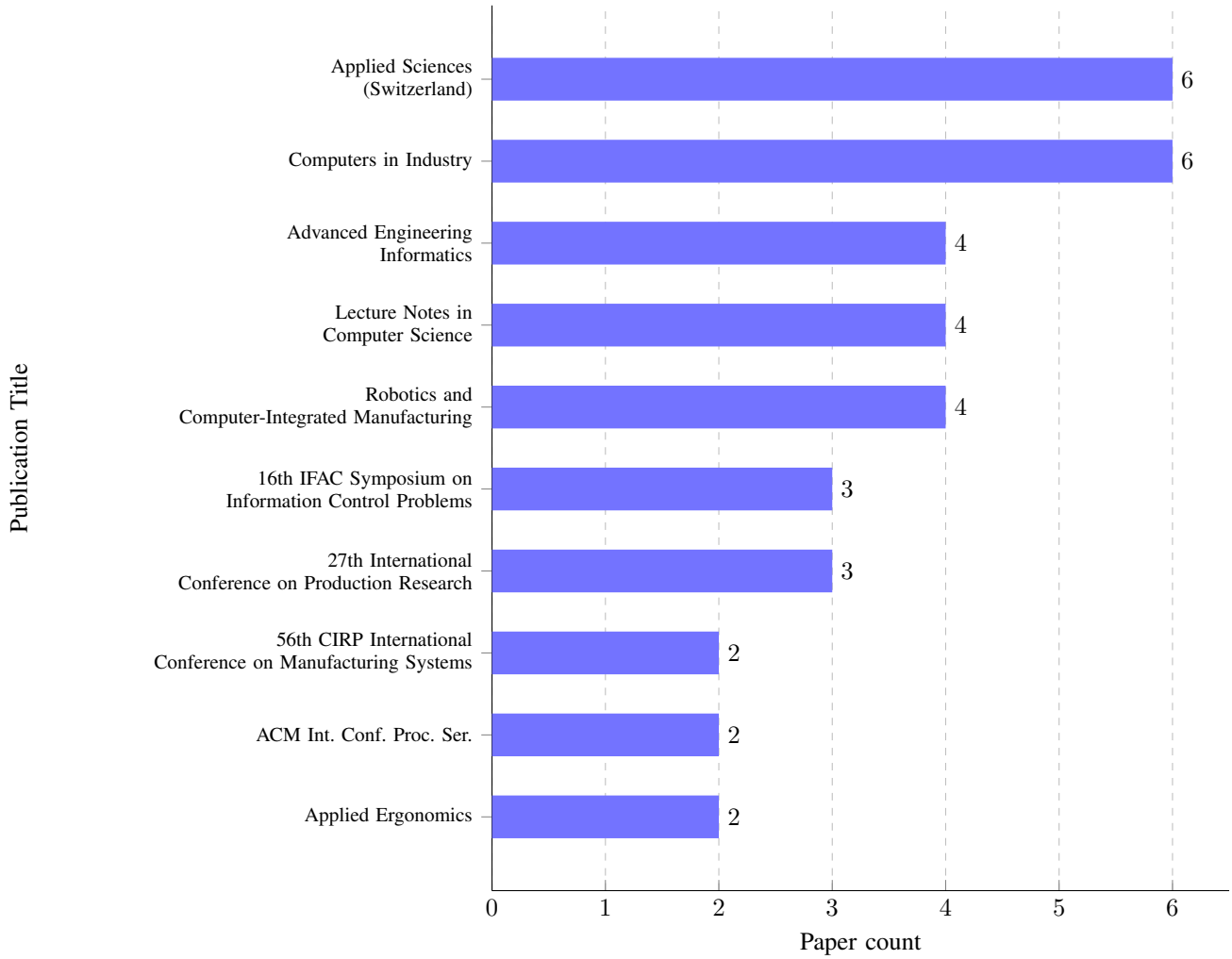


Fig. 7. Number of primary papers per publication venues.

phisticated spatial understanding [8], [14], [68], [70], [71]. The combination of real-world anchoring with immersive virtual elements makes MR particularly effective for collaborative troubleshooting and complex spatial tasks.

The chronological and application-based distribution patterns reflect both technological evolution and the maturing understanding of which XR technology best suits specific maintenance contexts. This suggests that industrial maintenance is shifting toward a task-specific application of XR technologies, rather than a one-size-fits-all approach.

1) Hardware trends: The analysis of hardware platforms across the reviewed studies reveals clear preferences and evolutionary trends in the adoption of XR devices for industrial maintenance applications. Table I presents the distribution of hardware platforms identified in the primary studies.

E. Head-Mounted Displays (HMDs)

Head-mounted displays represent the most frequently utilized hardware category in industrial maintenance applications, with Microsoft HoloLens being the dominant platform, mentioned in 18 studies. First appearing in our dataset in 2017 [21], HoloLens quickly established itself as the preferred option for industrial AR implementations, particularly for hands-free maintenance guidance [33], [60], [62], [66]. The prevalence of HoloLens can be attributed to its enterprise focus, robust tracking capabilities, and integrated computing power, making it suitable for industrial environments.

For VR applications, the HTC Vive headset emerged as the primary choice in four studies [12], [48], [64], [77], with limited mentions of other platforms, such as the Valve Index [8] and Oculus/Meta devices [79]. Study [8] represents an interesting case where a Valve Index headset was utilized for remote teleoperation of a mobile manipulator robot in cleanroom environments, highlighting the adoption of high-end consumer VR hardware for specialized industrial applications.

The temporal analysis of HMD adoption shows a distinct evolution in hardware preferences. Early studies (2014-2016) primarily employed non-HMD approaches, such as projector-based systems [22]. From 2017 onward, we have observed a clear shift toward head-mounted solutions, characterized by increasing sophistication in tracking and display capabilities. Studies [2], [3], [4], [75] illustrate the recent trend toward using HMDs to reduce task completion time, mental workload, and

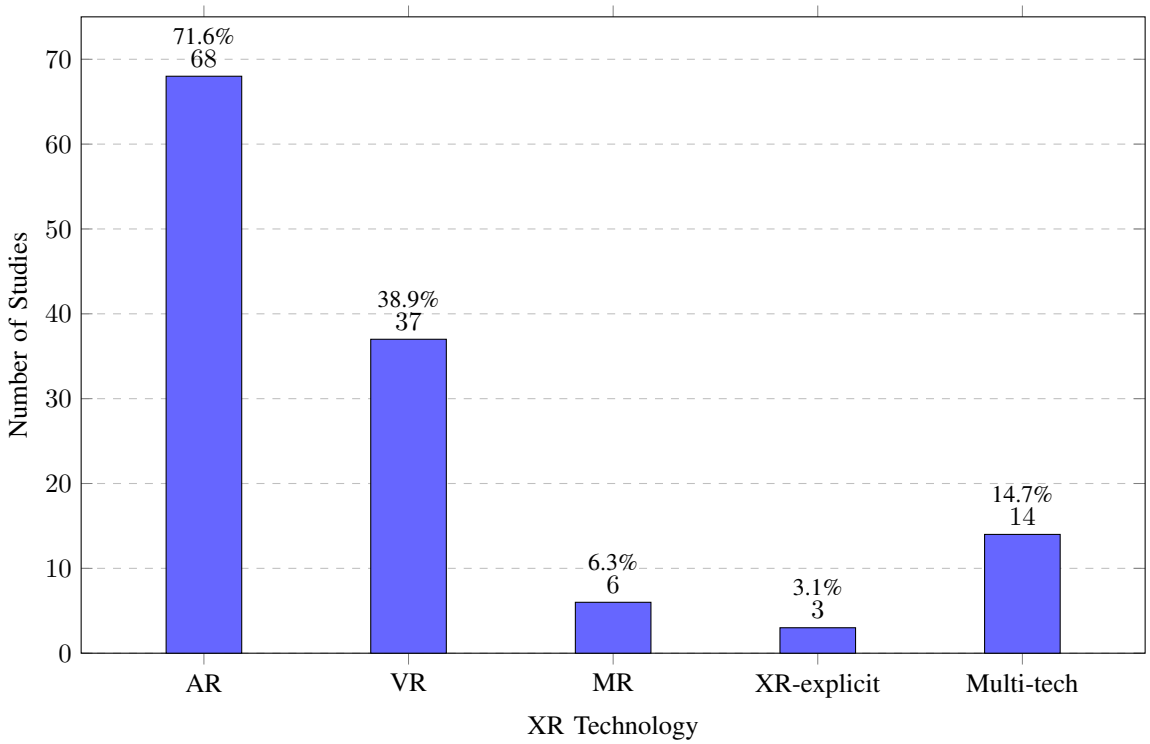


Fig. 8. Distribution of XR technologies across 95 primary studies. Percentages sum to over 100% as 14.7% of studies employed multiple technologies simultaneously.

improve cognitive processing for complex tasks.

F. Handheld Devices

Tablets (9 studies) and smartphones (6 studies) constitute the second most common hardware category, particularly prevalent in AR implementations where hands-free operation is not critical. Study [31] compared the effectiveness of AR on smart glasses (Vuzix M100) with AR on tablets, finding that smart glasses performed better in metrics such as completion time (-13% vs. -3% compared to paper) and scored higher in usability ratings and operator reliance. At the same time, both AR solutions demonstrated similar benefits in reducing errors compared to paper-based instructions. Studies [45], [69] illustrate the use of these devices, highlighting their advantages in terms of accessibility and user familiarity.

The temporal distribution shows consistent use of handheld devices throughout the review period, with notable spikes in 2016, 2018, and 2024. This persistent adoption of handheld solutions, even as HMDs became more sophisticated, suggests that different form factors serve complementary roles in industrial maintenance. Study [33] explicitly explored this complementarity by proposing a multi-device assistive system that leverages the strengths of both HMDs and handheld devices across different maintenance phases.

G. Other Hardware Platforms

Several supplementary hardware technologies appear in the reviewed studies, with haptic devices (6 studies) being the most prominent. Early implementations [1] utilized dedicated haptic systems like the Phantom Desktop, while more recent studies [39], [48] integrate haptic feedback with VR environments to enhance training realism.

Projector-based AR systems (4 studies) represent an alternative approach that avoids the need for personal displays entirely. Study [37] presents a portable camera-projector system for telemaintenance that enables spatially-registered annotations projected directly onto the work environment. This approach offers advantages in collaborative settings where multiple technicians need to view the same augmented workspace simultaneously.

Motion tracking technologies like Leap Motion controllers (2 studies) appear in applications requiring precise hand interaction, particularly for VR-based training systems [48], [49], [50]. These technologies enable natural gesture-based interaction with virtual environments, enhancing the realism of maintenance simulations.

H. Hardware Evolution and Limitations

The evolution of hardware platforms over the review period (2014-2025) reflects broader technological trends in the development of XR. Early studies employed more experimental or custom hardware configurations, whereas recent research has increasingly leveraged commercial off-the-shelf solutions with standardized development environments.

Several studies [5], [7], [15], [55], [59], [60], [62], [72] explicitly discuss hardware limitations as significant challenges, particularly regarding the limited field of view, ergonomic issues with prolonged use, and environmental constraints in industrial settings. Study [31] notes specific issues with smart

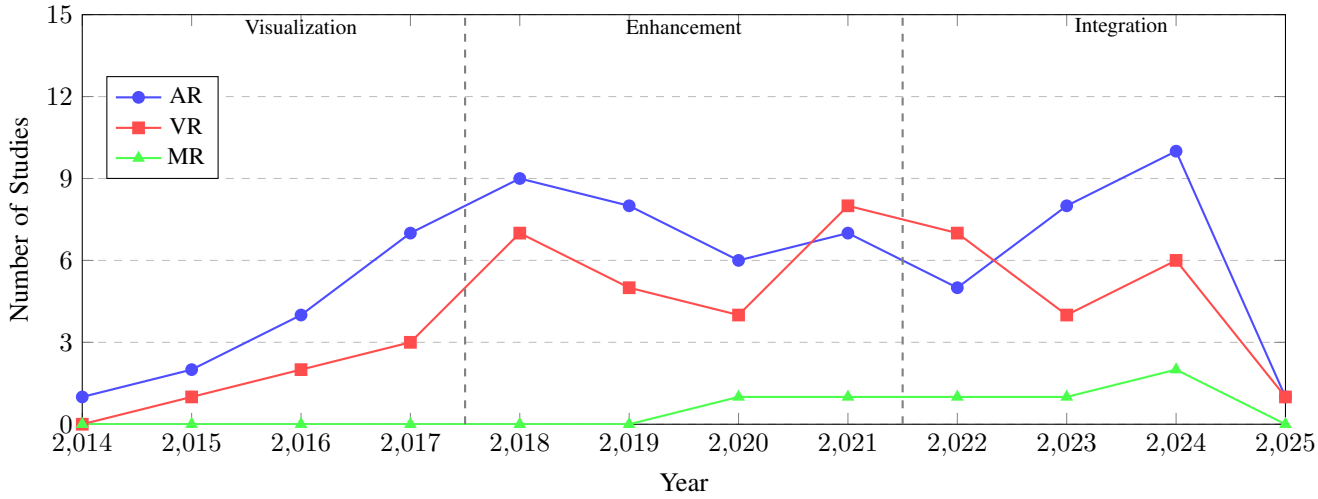


Fig. 9. Temporal evolution of XR technologies in industrial maintenance (2014-2025), showing three distinct evolutionary phases.

glasses, including comfort problems and a restricted screen size, which makes augmentations difficult to distinguish. This can cause users to close one eye to see correctly and requires constant visual focus switching, leading to visual strain.

The hardware trends suggest an industry in transition, with increasing standardization around enterprise-focused platforms, such as HoloLens for AR applications and HTC Vive for VR applications, while maintaining a pragmatic approach to device selection based on specific maintenance contexts and requirements.

1) Software and development: The software platforms, development frameworks, and implementation approaches used across the reviewed studies reveal essential patterns in how XR technologies are being integrated into industrial maintenance applications. This subsection examines the prominent software trends identified in the literature (Table II).

I. Development Platforms and Game Engines

Unity (identified in 18 studies) emerges as the dominant development platform across all XR technology types, serving as the foundation for applications in AR, VR, and MR. Its prevalence can be attributed to its cross-platform capabilities, extensive asset ecosystem, and relatively accessible learning curve. Study [1] used Unity to develop a haptic virtual reality platform for assembly tasks, while study [46] employed it to create an AR application for providing adaptive instructions for machine operation tasks. Study [12] also exemplifies the typical Unity-based development approach, utilizing the platform to create a VR system with AR simulation and gaze tracking for prototyping industrial maintenance solutions.

Unreal Engine appears in significantly fewer studies (only 1 explicit mention), despite its reputation for high-fidelity graphics and robust performance. Study [64] uses Unreal Engine 4 to develop a VR application for power plant maintenance training, suggesting that this engine may be preferred for applications requiring advanced visual fidelity.

The preference for Unity across the majority of studies indicates a maturation of the development ecosystem, where researchers and developers increasingly leverage established platforms rather than developing custom solutions from scratch. This standardization facilitates more rapid prototyping and reduces technical barriers to XR implementation in industrial contexts.

J. AR/VR Software Development Kits (SDKs)

Vuforia emerges as the most widely utilized AR SDK (6 studies). Studies [15], [21], [79] utilized Vuforia in combination with Unity to implement visual tracking and registration for maintenance tasks. The prevalence of Vuforia aligns with its established position in the AR development ecosystem, extensive documentation, and relatively stable performance across different deployment scenarios.

Other SDKs appear less frequently but highlight the diversification of development approaches. OpenXR and SteamVR in [81] represent newer standards-based approaches to VR development, while Microsoft's Mixed Reality Toolkit (MRTK) appears in [75] focused on HoloLens applications. Study [81] specifically mentions a combination of OpenXR, XRInteractionToolkit, and SteamVR packages within Unity, illustrating the trend toward utilizing multiple complementary frameworks within a single development environment.

K. Programming Languages and Technical Approaches

Specific programming languages receive relatively few explicit mentions in the methodology sections of the reviewed studies. Python (3 studies) is the most frequently cited, particularly in studies involving machine learning components or data processing pipelines [69], [74], [75]. C++ and C# are also used as programming languages with lower frequency, 1 explicit mention each [1], [34], with C# specifically appearing in a study focused on XR-based serious game development for Compressed Air System (CAS) maintenance training, where it was paired with the Unity development platform to create an immersive learning experience that incorporated 3D models and the XR Interaction Toolkit [34]. Study [1] presents a more complex development approach, combining OpenHaptics, PhysX, and OpenGL/GLUT libraries to create a haptic

TABLE I. OVERVIEW OF HARDWARE PLATFORMS USED IN XR-BASED INDUSTRIAL MAINTENANCE APPLICATIONS (2014–2025).

Category	Dominant Platforms (Count)	Primary Application Patterns
Head-Mounted Displays (HMDs)	• Microsoft HoloLens (18) • HTC Vive (4) • Valve Index (1) • Oculus/Meta devices (1)	• **HoloLens:** Dominant standard for hands-free AR maintenance guidance. • **HTC/Oculus:** Primarily used for immersive VR training simulations. • **Valve:** Specific usage for remote teleoperation in cleanrooms.
Handheld Devices	• Tablets (9) • Smartphones (6)	• **Tablets/Smartphones:** AR implementations where hands-free operation is not critical like inspection tasks requiring larger visual interfaces.
Supplementary Technologies	• Haptic devices (6) • Projector-based AR (4) • Motion tracking (2)	• **Haptics:** Integrated to enhance physical realism in VR training. • **Projectors:** Used for collaborative telemaintenance without wearable gear. • **Tracking:** Leap Motion used for natural gesture-based interactions.

VR platform for assembly planning and assessment. Such multi-library implementations are less common in more recent studies, suggesting a shift toward more integrated development environments that encapsulate these functionalities.

L. Content Authoring Approaches

A notable trend across the reviewed studies is the increasing focus on content authoring approaches that simplify the creation of XR maintenance applications without requiring extensive programming expertise. Studies [21], [23], [56], [58], [69], [72] present frameworks or methodologies specifically designed to enable non-technical domain experts to create and manage XR content for maintenance tasks.

Study [69] exemplifies this trend with an automatic content creation system for AR maintenance applications that extracts information from electronic manuals to generate AR-ready content accessible via QR codes. Similarly, the study [56] presents the INTERVALES framework that simplifies AR/VR content authoring for industrial applications by providing a modular architecture that supports both IT developers and domain experts.

This evolution toward simplified content creation constitutes a significant advancement in the practical deployment of XR in industrial maintenance, addressing a primary obstacle to implementation identified in earlier research — the high technical expertise traditionally required to develop and maintain XR applications.

M. Integration with Industrial Systems

The incorporation of XR applications into established industrial systems and information infrastructure represents an emerging focus area in more recent studies. Several papers [33], [38], [41], [45], [53] discuss approaches for connecting XR applications with industrial IoT systems, databases, and enterprise asset management systems.

Study [38] presents an integrated approach combining IoT, machine learning, and AR within a comprehensive predictive maintenance framework. Similarly, study [79] demonstrates a unified platform integrating IoT, analytical processing, AR, and VR technologies to support equipment surveillance and maintenance. These integrated approaches represent a maturation of the field, moving beyond isolated XR implementations toward more holistic solutions that leverage multiple complementary technologies.

The software and development trends reflect an industry transitioning from experimental proof-of-concept implementations toward more standardized and integrated approaches, with increasing emphasis on reducing technical barriers to adoption and enhancing integration with existing industrial ecosystems.

1) Application areas: The analysis of primary studies reveals distinct patterns in the application of XR technologies across maintenance tasks and scenarios in industrial environments. This subsection categorizes and examines the primary application areas identified in the literature, as summarized in Fig. 10.

N. Procedural Guidance and Task Assistance

Providing step-by-step guidance for maintenance procedures was the most prevalent application area (41 studies, 43.1%), with Augmented Reality the predominant technology in this domain. AR's ability to overlay digital information directly onto physical equipment makes it particularly suited for procedural guidance.

Studies [4], [15], [22], [27], [43], [75] showcase AR's application for guiding users through maintenance procedures with visual overlays. Study [22] demonstrates an early implementation using a screen-based video see-through AR system for motorcycle engine maintenance, reporting a 38% improvement in completion time and 92.4% reduction in error rate compared to traditional paper manuals. Recently, study [75] developed the TAGGAR system, which focuses on providing AR task guidance for procedural maintenance tasks by rendering virtual hand animations that demonstrate required

TABLE II. SUMMARY OF SOFTWARE TRENDS AND DEVELOPMENT FRAMEWORKS

Category	Dominant Tech	Key Implementation Patterns	Refs.
Game Engines	Unity, Unreal Engine	• Unity is dominant (18 studies) due to cross-platform assets. • Unreal used specifically for high-fidelity needs. • Shift from custom engines to standardized platforms.	[8], [12], [15], [21], [27], [29], [34], [46], [48], [50], [57], [65], [74], [75], [77], [79], [80], [81], [64]
SDKs	Vuforia, OpenXR, MRTK	• Vuforia remains the standard for visual tracking. • Adoption of OpenXR/MRTK reflects a move toward open standards and specific HoloLens optimization.	[15], [21], [29], [57], [73], [75], [79]
Languages	Python, C#, C++	• Python: Used for ML and data pipelines. • C#: Primary language for Unity-based interaction. • C++: Usage declining; mostly for legacy or haptic libraries.	[1], [34], [69], [74], [75]
Authoring	Low-code Frameworks	• Focus on simplifying creation for non-technical experts. • Automated content generation from manual data (e.g., QR codes).	[21], [56], [69]
Integration	IoT, EAM, Cloud	• Move from isolated apps to holistic ecosystems. • Integration with live IoT data and predictive maintenance systems.	[38], [45], [79], [81]

actions. This enables AR-based guidance for tasks such as pressing buttons, twisting knobs, pulling objects, and picking up/placing items.

The presentation modalities for procedural guidance have evolved considerably over the review period. Early studies primarily used text and simple 2D graphics, whereas more recent research has explored 3D animations [4], standardized graphical symbols [25], [26], [61], and context-aware adaptive interfaces [15], [21], [55]. Study [36] explicitly compares three different AR presentation modes for maintenance instructions in blind areas, finding that in-situ 3D presentations with X-ray rendering outperform 3D side-by-side in terms of completion time.

Several studies [21], [46], [47], [65], [76], [78] focus on adapting the level of guidance to match user expertise and task complexity. Study [47] proposes a system that combines collaborative robotics and AR, designed to accommodate workers with different experience levels through robotic assistance and AR guidance in car panel fitting operations. Study [76] introduces an approach to developing Augmented Reality-based Adaptive Assistance Systems (ARAAS) that adapt content to operator characteristics.

O. Training and Skill Development

Training is the second most common application area (mentioned in 32 studies, 33.6%), serving as a cornerstone use case for XR technologies in industrial maintenance. The engaging and interactive characteristics of XR environments offer a safe, controlled space for skill development without the risks or resource constraints of real-world training scenarios.

Virtual Reality dominates the training domain, with studies [10], [30], [35], [63], [64], [77] showcasing comprehensive VR-based training systems for various maintenance procedures. Study [35] demonstrates how VR training enhances performance in intricate industrial assembly tasks compared to traditional methods, showing improved performance and training permanency. Similarly, study [77] compares VR-based training with traditional paper-based approaches for wiring electric motors, reporting a 10% improvement in task completion time.

Augmented Reality training applications focus more on in-situ guidance and real-time skill development. Study [29] presents an AR mobile application for training on CNC lathe maintenance and repair, highlighting improved student motivation and learning outcomes. The hybrid approach of using AR during actual maintenance tasks blurs the line between training and operational support, as illustrated in study [17], which compares AR, VR, and video-based training for manual assembly tasks.

More recent studies [80], [81] demonstrate an evolution toward more sophisticated training approaches, incorporating game-based elements, adaptive difficulty, and integration with real equipment data. Study [80] introduces an XR-based serious game for compressed air system maintenance, while study [81] presents a modular framework for developing VR environments to enhance safety training during human-robot collaboration scenarios.

P. Assembly and Disassembly Support

Assembly and disassembly tasks appear as specific focus areas within the broader maintenance domain, with assembly operations mentioned in 24 studies (25.2%) and disassembly processes in 5 studies (5.2%). These complementary processes represent fundamental maintenance activities that benefit significantly from XR guidance.

AR dominates dis/assembly support applications, as demonstrated in studies [4], [16], [24], [41], [57], [59]. Study [41] introduces an AR solution that facilitates the assembly of multi-stage rotors used in turbine-based combined-cycle engines. Study [16] presents an AR-assisted framework for product disassembly (ARDIS) to improve efficiency in product maintenance and remanufacturing processes. Similarly, study [24] applies Design for Disassembly (DfD) principles with AR visualization to optimize the disassembly process for industrial components.

VR applications in this domain focus more on planning, verification and simulation aspects of assembly/disassembly. Study [1] presents a haptic VR platform for planning, performing and assessing virtual assemblies, providing force feedback and realistic part behavior. Study [67] introduces a multimodal system for verifying assembly procedures and product design in VR using CAD models and haptic feedback.

Mixed Reality applications for assembly tasks appear in studies [68], [70], with a particular focus on remote collaboration scenarios. Study [70] presents an adaptive MR remote collaboration system that provides clear, detailed, and spatially correct instructions for assembly tasks, demonstrating improved performance and reduced workload compared to a typical remote collaboration interface called RECI.

Q. Remote Assistance and Collaboration

Remote expert support and collaborative maintenance emerged as a significant application area (22 studies, 23.1%), reflecting the growing importance of distributed work and knowledge sharing in industrial settings. All three major XR technology types (AR, VR, and MR) are well-represented in this domain, each offering distinct advantages.

AR-based remote assistance systems are demonstrated in studies [7], [9], [37], [66], with study [37] presenting a projector-based AR system for telemaintenance that enables bidirectional spatial communication between remote experts and on-site technicians. AR's ability to provide spatially-anchored guidance in the real environment makes it particularly effective for remote support scenarios. VR-based collaborative environments have been explored in studies [11], [13], [14], [32], focusing on shared virtual spaces for maintenance planning and knowledge exchange. Study [32] examines the application of collaborative VR for reviewing maintenance documentation and assessing risk, highlighting numerous benefits for globally distributed teams.

The most sophisticated remote collaboration approaches utilize Mixed Reality, as shown in studies [68], [70], [71]. Study [70] presents an MR remote collaboration system (AGICA) that supports adaptive instruction generation based on context awareness, while study [71] introduces a novel MR system that enables remote experts to create 3D models for annotating spatial areas with visual notifications about interactions.

R. Inspection and Diagnostic Support

Inspection tasks (5 studies, 5.2%) and diagnostic/troubleshooting support (3 studies, 3.1%) represent more specialized application areas where XR technologies augment human perceptual and analytical capabilities. These applications often combine XR with additional sensing modalities or analytical tools.

Study [20] presents an innovative system (AROD) that combines AR with an electronic nose (e-nose) for odor detection during maintenance, displaying relevant information to detected odors via a HoloLens headset. Similarly, study [51] introduces MANTRA, a novel AR system that combines AR and infrared thermography for industrial maintenance, automatically aligning virtual information and temperature data with 3D objects.

In the inspection domain, study [41] presents an AR-based system for guiding the assembly of multi-stage rotors, incorporating measurement equipment to provide real-time visual assembly guidance tailored to precision requirements. Study [76] applies an adaptive AR assistance system specifically to landing gear inspection operations, highlighting areas to measure and displaying corresponding tolerance values, resulting in reduced execution time and improved information access.

Recent studies indicate an increasing integration of XR with other Industry 4.0 technologies to enhance diagnostic capabilities. Study [53] presents a methodology for automatic segmentation and analysis of anomalies in solar PV modules using unsupervised sensing algorithms coupled with 3D AR visualization. Study [79] demonstrates an integrated system combining IoT, data analysis, AR and VR for equipment monitoring and maintenance, enhancing predictive maintenance capabilities and operational efficiency.

The literature identifies diverse and evolving XR use cases in industrial maintenance. While specific applications, such as training and procedural guidance, are well-established across the review period, newer applications, including integrated diagnostic systems and adaptive assistance frameworks, represent emerging trends that leverage XR's capabilities in increasingly sophisticated ways.

S. Advantages of Extended Reality in Industrial Maintenance

The analysis of primary studies reveals five major categories of advantages associated with XR technologies in industrial maintenance contexts, as summarized in Table III. These benefits collectively demonstrate the transformative potential of these technologies across various aspects of maintenance operations.

1) Efficiency and performance improvement: Improved operational efficiency represents one of the most frequently cited advantages of XR technologies. AR applications consistently demonstrate significant time savings in maintenance execution. Study [22] reports a 38% improvement in completion time for motorcycle engine maintenance tasks compared to traditional paper manuals. Study [76] presents a methodology for designing Augmented Reality-based Adaptive Assistance Systems (ARAAS). The system supports inspection operations of landing gear components during maintenance, repair, and overhaul (MRO) activities, resulting in a 10% reduction in total inspection time.

The efficiency gains are pronounced for complex maintenance tasks as well. Study [2] demonstrates that AR instructions significantly reduce task completion time, particularly for complex maintenance activities, whereas they have a smaller impact on simpler tasks. Similarly, study [3] shows that task completion time is reduced with AR-based instruction methods compared to paper-based methods, especially for complex tasks involving a piston pump system. This finding suggests that escalating benefits occur as task complexity increases — a valuable characteristic in industrial environments, where complex procedures pose the most significant challenges.

The cost implications of these efficiency improvements stem from decreased downtime [7], [11], reduced travel requirements for experts [7], [9], [11], and more efficient use of

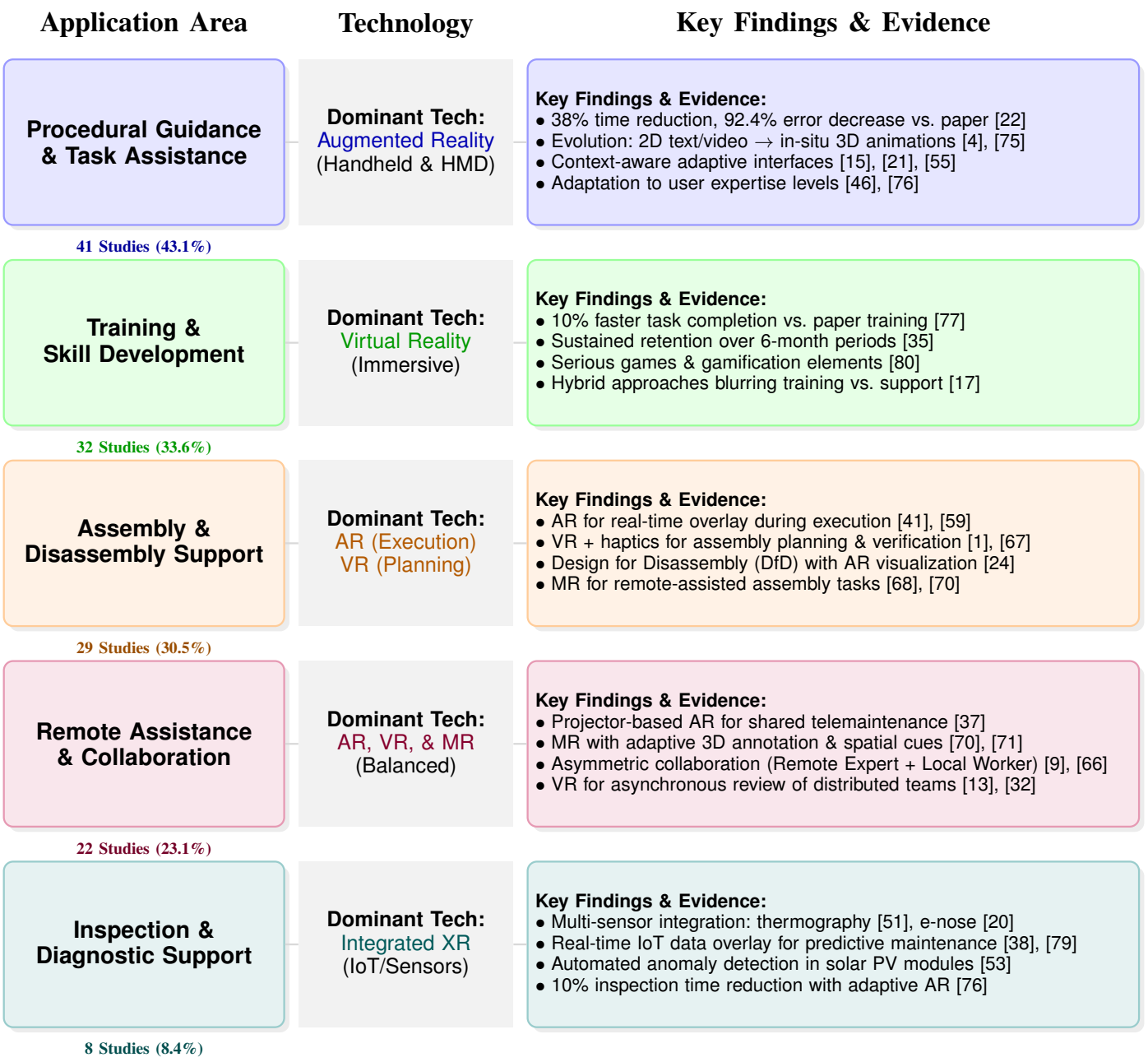


Fig. 10. Taxonomy of primary application areas in XR industrial maintenance, mapping study frequency, dominant technologies, and key empirical findings with supporting references.

personnel [2], [9], [11]. Notably, study [45] identifies cost reduction as a direct benefit of AR-based maintenance assistance. In contrast, study [64] highlights the economic advantages of VR training over physical equipment use, including reduced costs, faster training delivery, and the elimination of expensive physical resources.

2) Quality enhancement and error reduction: Error reduction emerges as a significant advantage in reviewed studies. The visual guidance and procedural structure provided by XR applications contribute to fewer errors and more consistent performance. Study [22] demonstrates a remarkable 92.4% reduction in error rates when using AR for maintenance tasks compared to paper-based instructions. Similarly, study [31] found that AR guidance resulted in fewer errors compared to paper instructions during maintenance tasks.

VR applications show comparable benefits in training contexts. Study [35] reports that VR-based training for industrial assembly tasks led to improved performance and fewer errors during subsequent real-world execution. The immersive environment enables repeated practice without risk, improving proficiency before physical interaction with equipment.

MR offers unique advantages for quality in collaborative scenarios. Study [70] presents an MR system that reduced errors by providing spatially correct instructions for assembly

TABLE III. SUMMARY OF KEY ADVANTAGES AND BENEFITS OF XR IN INDUSTRIAL MAINTENANCE

Category	Key Benefits & Empirical Evidence	Refs.
Efficiency & Performance	• **Time Reduction:** Significant savings in execution time (e.g., 38% improvement in engine maintenance ; 10% in inspection tasks). • **Complexity Scaling:** Benefits escalate as task complexity increases (e.g., piston pump systems) [2], [3]. • **Cost Impact:** Reduced downtime, decreased expert travel, and elimination of physical training assets [45], [64].	[2], [3], [7]
Quality & Error Reduction	• **Error Mitigation:** Visual overlays prevent mistakes. Reported 92.4% error reduction vs. paper manuals [22]. • **Precision:** Spatially correct 3D instructions prevent misalignment in assembly tasks [70]. • **Training Transfer:** Virtual practice leads to fewer errors during subsequent real-world execution [35].	[22], [31], [35], [70]
Learning & Knowledge Transfer	• **Risk-Free Practice:** VR enables "learning-by-doing" without equipment damage [10]. • **Retention:** Improved training permanency observed over 6-month periods [35]. • **Just-in-Time:** AR provides contextual knowledge directly at the point of need [42].	[10], [32], [35], [42], [77]
Human-Centric Advantages	• **Cognitive Load:** Reduces attention switching between manuals and tasks; aids processing in complex procedures [2], [16], [54]. • **Adaptability:** Content adapts to user skill levels (novice vs. expert) and preferences [42], [46], [76]. • **Ergonomics:** Hands-free operation via HMDs and voice assistants reduces physical strain [33], [62], [73].	[25], [42], [46], [54], [73], [76]
Collaboration & Safety	• **Remote Support:** Expert guidance for distributed teams; asynchronous collaboration on virtual prototypes [9], [13]. • **Hazard Awareness:** Visualization of safety zones (e.g., robot workspaces) and risk mitigation planning [19], [81].	[9], [13], [19], [28], [81]

Note: Where multiple references appear in a single cell, inline citations within the Key Benefits/Barriers column indicate the specific claim each reference supports. References listed in the Refs. column without inline attribution provide additional evidence for the category as a whole.

tasks. The visualization of 3D elements in their proper spatial context helps prevent misalignment and assembly errors.

3) Enhanced learning and knowledge transfer: XR technologies demonstrate significant benefits for maintenance training and knowledge transfer. VR creates safe environments for practice, eliminating risk to equipment and personnel. Study [10] explains that VR provides immersive simulations that enable learning-by-doing in risk-free settings, potentially shortening learning periods. Study [77] offers empirical evidence that VR-based training for wiring electric motors led to 10% faster task completion than traditional training.

AR facilitates learning directly within the work context, enabling just-in-time knowledge acquisition. Study [42] demonstrates that AR guidance enhances skill development by providing personalized instruction tailored to individual skill levels. Knowledge retention represents another advantage, as a study [35] reports improved training permanence with VR-based approaches, demonstrating sustained performance improvements over six months.

These XR learning benefits extend to knowledge sharing across organizations. Study [32] demonstrates how VR enhances collaboration and spatial understanding for globally distributed teams performing maintenance, documentation review, and risk assessment.

4) Human-centric advantages: Human-centric benefits address cognitive enhancement, ergonomic improvements, and personalized experiences. These advantages directly impact operator well-being and performance in maintenance contexts.

Cognitive Load Optimization: XR technologies effectively modulate cognitive load during complex procedures. Study [2] demonstrates that AR instructions enhance cognitive processing, particularly for complex tasks, by presenting information in a more intuitive and spatially relevant manner. Study [54] illustrates how AR reduces cognitive load through automatic object detection and spatial guidance, providing situation-aware and dynamic task assistance. Traditional maintenance approaches pose significant cognitive challenges, as highlighted in Study [16], which notes that operators must constantly shift their attention from the task at hand to instruction manuals, making disassembly processes error-prone and time-consuming. XR solutions address this fundamental limitation by overlaying contextual information directly onto the work environment, eliminating the need for attention switching and maintaining operator focus on the primary task.

Adaptability to User Characteristics: Approximately one-third of studies highlight personalization features that accommodate varying skill levels and learning styles. Study [76] presents an AR-based Adaptive Assistance System methodology that adapts content to user expertise level, adjusts

information presentation based on preferences, and provides personalized assistance experiences. Study [42] demonstrates that AR can enhance the accessibility of manufacturing jobs by enabling personalized delivery of task information tailored to workers' preferences. Study [46] an adaptive framework for generating Augmented Reality (AR) instructions for machine operation, tailored to individual operators' skill levels. The system retrieves task data and worker profiles to provide personalized AR guidance, thereby improving productivity, safety, and worker satisfaction.

Ergonomic Enhancement: XR technologies contribute to improved physical ergonomics in many studies. Head-mounted AR displays enable hands-free operation, as demonstrated in study [33], allowing technicians to access information while maintaining physical engagement with equipment. Study [73] provides objective measurements showing AR instructions can reduce physical strain in certain muscle groups, particularly for simple maintenance tasks.

Improved Information Design: Enhanced information presentation is another human-centric advantage. Study [25] presents a methodology for converting technical documentation into AR-based manuals, optimizing text usage and converting textual instructions into visual elements. Study [62] shows how AR voice assistants can facilitate hands-free interaction during manual tasks, improving both ergonomics and information accessibility.

5) Collaboration and safety: Improved collaboration is reported in more than one-third of studies, with a particular emphasis on remote expert support. Study [9] demonstrates how AR assists service technicians with remote expert guidance for problems of varying difficulty. Study [13] illustrates how the COVE-VR platform facilitates asynchronous collaboration among globally distributed departments, enabling interaction with virtual prototypes in realistic settings.

Safety benefits manifest through risk-free training environments and improved hazard awareness. Study [28] proposes a human-centered conceptual model that integrates Augmented Reality (AR) and Dynamic Digital Models (DDM) to improve training and mitigate occupational risks in industrial contexts, enabling real-time risk management support for occupational safety. Study [81] presents a template concept for creating and modifying VR environments more efficiently in industrial contexts, specifically focusing on safety training for human-robot collaboration in industrial assembly. The approach emphasizes visualizing safety zones and robot workspaces while helping operators understand safety protocols. Study [19] presents a methodological proposal called SMART (Sistema Modular de Acceso Rápido Temporal) for implementing Virtual Reality (VR) and Augmented Reality (AR) systems to train engineering students in industrial maintenance and safety skills without exposing them to real risks. The methodology consists of four phases: Requirements, Architecture, Construction, and Evolution. The primary tasks involve utilizing VR for risk identification, assessment, and mitigation planning in industrial settings.

6) Evolution of advantages over time: The advantages of XR technologies have evolved significantly over the 2014-2025 review period, reflecting both technological advancements and a deeper understanding of their implementation.

Early studies (2014-2017) focused primarily on fundamental efficiency improvements and basic visualization capabilities. Study [22] exemplifies this phase, emphasizing the reduction of completion time and error elimination through AR visualization. Advantages described in this period tended to be directly measurable, such as time savings and error counts.

Middle-period research (2018-2021) expanded to include more sophisticated cognitive and learning benefits. Studies [10], [35], [47] demonstrate greater attention to knowledge transfer, cognitive support, and adaptive capabilities. The advantages described become more nuanced, addressing not just faster task completion but also enhanced understanding and knowledge retention.

Recent studies (2022-2025) show a shift toward personalization, human-centric design, and integration with other technologies. Study [75] demonstrates this trend by presenting an AR system that automatically generates task guidance using vision-language models, a level of integration with AI not seen in earlier studies. Study [70] showcases advanced MR applications that adapt to contextual awareness, while study [2] provides empirical evidence for AR's differential benefits based on task complexity.

This evolution reflects the maturation of XR from novel technical solutions to sophisticated sociotechnical systems that actively adapt to human needs and organizational contexts, suggesting continued development toward increasingly intelligent, integrated, and human-centered XR applications for industrial maintenance.

T. Limitations and Challenges of Extended Reality in Industrial Maintenance

The implementation of XR technologies in industrial maintenance faces several significant barriers that must be addressed to realize their full potential. Analysis of the primary studies reveals four major categories of limitations and challenges that currently impede wider adoption of these technologies.

1) Technical limitations: Hardware constraints represent the most frequently reported technical limitation. Head-mounted displays (HMDs) are particularly criticized for their limited field of view, which restricts the user's peripheral vision during maintenance tasks. Study [7] identifies this narrow field of view as problematic for guiding technicians during on-site repair tasks and providing remote experts with AR views of machinery for analysis. Battery life limitations also affect deployment feasibility, with study [15] highlighting the need for extended battery life to support all-day use in industrial settings.

Tracking and registration accuracy emerge as another significant technical challenge. Study [54] reports image distortion during rapid motion, which affects the reliability of object detection, and notes that current AR devices struggle with reflective or transparent objects common in industrial equipment. Registration misalignment, where virtual elements do not precisely align with physical components, is highlighted in studies [4], [36], [52], [59] as particularly problematic for maintenance tasks requiring high precision.

Network performance and computational requirements present additional technical barriers. Study [70] identifies

limitations in network connectivity with current Wi-Fi technology as a constraint for data-intensive XR applications. Study [44] demonstrates these connectivity challenges in real-world field deployments, where poor network performance forced researchers to use dedicated 4G hotspots paired with HoloLens 2 devices to address connectivity issues, though even this solution proved insufficient for performance-intensive features that require high-bandwidth connections. Computational limitations identified across studies include processing complexity and performance constraints when handling large 3D CAD models [13], jitter in pose estimation algorithms [51], challenges in data management and scene reconstruction [71], and insufficient computational resources for complex scenarios requiring real-time processing and rendering.

2) Implementation and adoption barriers: Organizational implementation challenges appear in 42% of studies, spanning integration, acceptance, and resource constraints. The high implementation cost represents a significant barrier, particularly for small and medium enterprises. Studies [34], [64], [80] explicitly identify substantial initial investment requirements as a limitation, while study [29] points to financial resource constraints in adapting XR for technical training contexts.

User acceptance and adaptation difficulties are prominently reported in studies [9], [26], [35], [52], [59], [79]. Study [26] reports the need for user adaptation to symbol-based instructions in its proposed method, aiming to reduce the amount of text in technical documentation. Study [9] explicitly notes that resistance is present among technicians with lower technology-affinity scores. This reluctance manifests as skepticism about the accuracy and trustworthiness of information, with study [6] identifying potential trust issues in information provided through XR interfaces. Study [62] reports user frustration with speech recognition errors during maintenance tasks, highlighting the gap between expected and actual performance in real-world conditions.

Content-creation complexity poses a persistent challenge in implementation. Study [26] identifies the complexity of AR authoring as a significant barrier due to difficulties in creating scalable solutions for different types of maintenance manuals. Studies [47], [52] report the laborious model preparation required for effective AR implementation, thereby imposing a high threshold for initial deployment.

3) Human factors challenges: Cognitive and physical ergonomic issues are identified in 35% of studies. Extended use of HMDs can cause physical discomfort, with study [73] providing objective measurements showing increased strain in certain neck and shoulder muscles due to the weight and required posture of AR headsets. Study [3] reports the potential for cognitive overload due to focus rivalry between real and virtual information, particularly for complex tasks.

User skill variability presents challenges for designing appropriate XR interfaces. Study [40] identifies difficulties in providing proper guidance for users of varying skill levels, while study [3] notes a learning curve for novel AR interaction techniques. Study [55] reports initial challenges with gesture controls for users unfamiliar with AR interfaces.

Environmental considerations further complicate human factors design. Study [36] identifies implementation challenges related to varying lighting conditions, which affect tracking reliability. Study [20] notes the influence of humidity and temperature on sensing capabilities, while study [8] reports difficulties in detecting reflective or transparent objects when using Microsoft HoloLens smart glasses.

4) Methodological and evaluation challenges: Assessment limitations are also reported in the primary studies, particularly regarding the ecological validity of laboratory evaluations. Study [66] notes the difficulty in replicating real-world conditions in controlled tests and identifies challenges in balancing safety concerns with the requirements of real-world testing.

Many studies rely on limited sample sizes or non-representative participant groups. Study [76] acknowledges a key limitation in having only three inspectors test their ARAAS methodology, which potentially limits the generalizability of the results. Similarly, study [3] notes the limited generalizability of findings due to participant characteristics and the specific AR interface tested. Study [12] reports limited generalizability due to its focus on elevator maintenance, while presenting the xR Safety Kit framework for AR-based industrial maintenance solutions.

Long-term effectiveness evaluation remains particularly challenging. Study [17] identifies the need for studies examining extended XR use rather than single-session experiments. In contrast, study [3] notes limitations in understanding the long-term effects of AR use on user performance and adaptation.

These limitations and challenges, summarized in Table IV, highlight the current state of XR technologies in industrial maintenance as promising but still maturing. Many studies conclude with recommendations for future research that specifically target these barriers, suggesting a collective recognition of the need to address these challenges and advance the field toward more effective and widely implemented XR solutions.

U. Evaluation Criteria of XR in Industrial Maintenance

The systematic evaluation of XR technologies in industrial maintenance employs diverse criteria and instruments to validate effectiveness across multiple dimensions. Analysis of the primary studies reveals a structured evaluation landscape organized into five primary categories, as illustrated in Fig. 11, each addressing distinct aspects of XR system performance in maintenance contexts.

1) Performance-based evaluation: Objective performance metrics represent the most prevalent approach. Task completion time and error rates each appear in 25 studies, reflecting industrial emphasis on quantifiable operational improvements. Study [22] demonstrates AR's meaningful impact, reporting a 38% reduction in completion time and a 92.4% decrease in errors for motorcycle engine maintenance compared to paper instructions. Study [77] shows similar benefits for VR training, with 10% faster task completion for electric motor wiring. The combined assessment of efficiency (time) and effectiveness (errors) provides robust evidence for XR impact, as demonstrated across diverse maintenance scenarios in Studies [18], [36], [44].

2) Cognitive load and workload assessment: Cognitive evaluation appears in 16 studies, with the NASA Task Load Index used in 10 studies. This standardized questionnaire measures mental, physical, and temporal demands, as well as

TABLE IV. SUMMARY OF BARRIERS, LIMITATIONS, AND CHALLENGES IMPEDING XR ADOPTION IN INDUSTRIAL MAINTENANCE

Category	Key Barriers & Challenges	Refs.
Technical Limitations	• **Hardware Constraints:** Limited Field of View (FOV) restricts peripheral vision; battery life insufficient for full shifts [7], [15]. • **Tracking & Registration:** Misalignment of virtual overlays; instability with rapid movement or reflective/transparent surfaces [4]. • **Connectivity:** Wi-Fi/Network latency hinders data-intensive tasks; heavy processing of CAD models causes jitter [13], [70].	[4], [7], [13], [15], [70], [71]
Implementation & Adoption	• **Cost:** High initial investment and resource requirements for SMEs and training contexts . • **User Acceptance:** Resistance due to low tech affinity; trust issues with data accuracy; frustration with interaction errors (e.g., speech) [6], [9]. • **Content Creation:** Laborious model preparation and complex authoring processes create high deployment thresholds .	[6], [9]
Human Factors & Ergonomics	• **Physical Strain:** Neck/shoulder fatigue from HMD weight and prolonged posture . • **Cognitive Load:** Focus rivalry between real/virtual elements; cognitive overload in complex tasks . • **Environmental Sensitivity:** Lighting variations, humidity, and temperature affect sensor reliability and tracking .	[3], [8]
Methodological & Evaluation	• **Ecological Validity:** Difficulty replicating real-world industrial conditions (safety, chaos) in lab settings [66]. • **Generalizability:** Reliance on small sample sizes (e.g., N=3) or non-representative participant groups [12]. • **Longitudinal Data:** Lack of studies examining long-term effects, adaptation, and sustained performance [3], [17].	[3], [12], [17]

Note: Where multiple references appear in a single cell, inline citations within the Key Benefits/Barriers column indicate the specific claim each reference supports. References listed in the Refs. column without inline attribution provide additional evidence for the category as a whole.

performance, effort, and frustration. Study [2] reveals differential cognitive impact across task complexity, showing significant workload reduction for complex maintenance procedures, whereas simpler tasks yield minimal benefit. Study [3] extends this by measuring EEG (refers to the use of electroencephalography (EEG) to study and measure brain activity related to cognitive processes) during AR-assisted pump maintenance, revealing neural correlates of reduced cognitive demand. The integration of subjective workload measures with objective performance metrics provides complementary perspectives, as exemplified by the comprehensive evaluation of AR plant maintenance systems in Study citep7.

3) Physiological measurement: Physiological evaluation appears in 7 studies and offers unique insights through biosensing technologies. Heart rate variability emerges as the most common measure [2], [10] and assesses cardiovascular responses and physical stress. Electrodermal activity provides a complementary assessment of stress in Studies [4], [10]. Eye tracking offers particular value for evaluating visual attention and information processing; Study [4] employed comprehensive gaze behavior analysis during AR-assisted assembly, and Study [12] used it to identify safety hazards in VR simulations. Despite the advantages of continuous, real-time assessment and capture of unconscious responses, the relatively low adoption rate likely reflects practical constraints, including specialized equipment requirements and demands for technical expertise.

4) Usability and user experience: Usability assessment appears in 28 studies, emphasizing that technical functionality requires intuitive, satisfying user interactions. The System Usability Scale (SUS) emerges as the most widely adopted standardized instrument in 10 studies, enabling comparison across different XR implementations [15], [17], [54], [55]. User experience evaluation extends beyond narrow usability to encompass satisfaction and acceptance, appearing in 33 studies. Study [5] evaluates integration into existing maintenance workflows, while Study [11] focuses on perceived usefulness for industrial tasks. Several studies employ domain-specific instruments, such as ISONORM [7], which are tailored to industrial contexts and ergonomic standards.

5) Specialized dimensions and multi-method approaches: Learning outcomes are explicitly evaluated in 13 studies, particularly for VR training applications. Study [35] assesses six-month knowledge retention following VR-based assembly training and demonstrates sustained improvements in performance. Economic evaluation appears in 8 studies, with Study [64] quantifying the cost advantages of VR training through reduced training time and resources and its cost-effectiveness as a training solution. Safety assessment in 5 studies and collaboration quality in 5 studies address specialized application contexts.

A comprehensive evaluation employing multiple assessment modalities is reported in 18 studies. Study [4] exemplifies this approach, integrating five dimensions: task completion time, error analysis, NASA-TLX, physiological measures (EDA and inter-beat interval), and usability (PSSUQ). Study [7] evaluates effectiveness, efficiency, satisfaction, situational awareness (SART), and usability (ISONORM). This multimethod strategy enables the identification of potential trade-

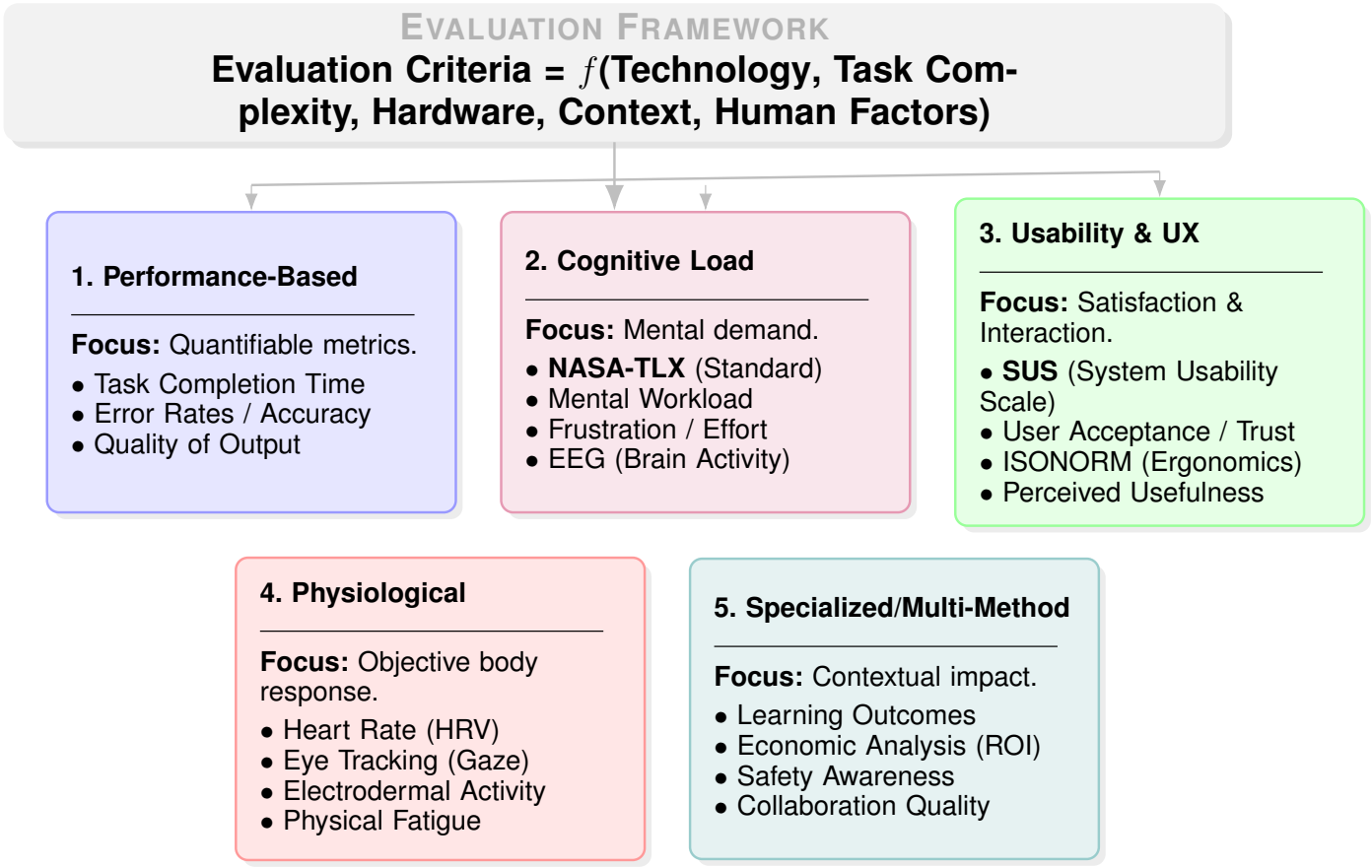


Fig. 11. Comprehensive evaluation framework for XR in industrial maintenance, illustrating the five primary assessment categories and the functional relationship defined in this review.

offs and provides both quantitative evidence and a nuanced understanding of user experiences.

6) Technology-specific patterns and framework: The relationship between evaluation approaches and implementation contexts reveals strategic patterns. Complex maintenance tasks require comprehensive evaluation, including cognitive load and physiological measures, as demonstrated in Studies [2], [3], [4]. Human-centric evaluation integration is featured in 42 studies, addressing cognitive support, ergonomics, adaptability, and user experience-critical factors for the adoption of industrial XR.

The comprehensive evaluation framework (see Fig. 11) illustrates the relationship: Evaluation Criteria = f(XR Technology, Task Complexity, Hardware, Application Area, Human-Centric Factors). This functional relationship informs the design of systematic evaluations, with criteria selection tailored to the type of technology (AR's operational focus suggests performance metrics; VR's training focus emphasizes learning outcomes), task complexity (complex scenarios justify using multiple methods), and human-centric factors (adaptive systems require assessing personalization effectiveness). As XR technologies advance and are adopted more widely, evaluation methods must also be developed to deliver strong evidence that supports confident industrial adoption decisions.

IV. Discussion

This systematic literature review synthesizes evidence from primary studies on extended reality (XR) technologies in industrial maintenance, revealing a rapidly maturing field transitioning from technology-centric implementations toward more sophisticated, human-centered systems.

A. Summary of Key Findings

The analysis reveals fundamental patterns in the adoption of XR for industrial maintenance. AR dominates current implementations (71.6%), primarily serving operational guidance and on-site assistance. VR accounts for 38.9% of studies, with strength in training and simulation contexts. MR represents only 6.3% but demonstrates emerging potential for collaborative maintenance scenarios. This distribution reflects pragmatic alignment between technology characteristics and maintenance task requirements.

Temporal analysis illuminates three distinct evolutionary phases. Early research (2014-2017) emphasized fundamental efficiency improvements through basic visualization. Middle-period studies (2018-2021) expanded to cognitive and learning benefits, incorporating adaptive capabilities and knowledge transfer mechanisms. Recent work (2022-2025) demonstrates AI integration, personalization, and context-aware systems with automatic content generation. This progression mirrors

the transitions from Maintenance 4.0 to Maintenance 5.0, where human-centricity emerges as a central design principle.

The prominence of human-centric factors across 44.2% of studies signals an evolving research emphasis. Temporal analysis across the three identified phases supports this interpretation: human-centric factors appeared in approximately 25% of studies during Phase 1 (2014–2017), increased to 42% during Phase 2 (2018–2021), and reached 58% during Phase 3 (2022–2025). This progressive increase substantiates the characterization of a paradigm shift toward human-centered XR design in industrial maintenance. Evaluation approaches have matured, with 18.9% of studies employing comprehensive, multi-method assessments that integrate performance metrics, cognitive load measures, physiological indicators, and user experience dimensions.

B. Theoretical Contributions

This review presents two frameworks that advance the understanding of XR effectiveness in maintenance contexts.

The Evaluation Framework expresses effectiveness as a functional relationship: Evaluation Criteria = f(XR Technology, Task Complexity, Hardware, Application Area, Human-Centric Factors). This framework guides systematic assessment by recognizing that evaluation criteria must align with the characteristics of technology, task demands, and human factors. Complex maintenance procedures requiring AR guidance with HoloLens 2 in field deployments necessitate a comprehensive evaluation encompassing performance metrics, NASA-TLX cognitive assessment, usability measures, and physiological indicators. VR training applications require the measurement of learning outcomes, knowledge retention, and transfer effectiveness.

The Technology-Task-Context Alignment Model proposes optimal pairings between XR modalities, maintenance tasks, and operational contexts. AR demonstrates superiority for guidance, inspection, and procedural support in on-site environments where real-world context is essential. VR excels in training, planning, and simulating hazardous procedures in controlled learning environments. MR facilitates collaborative problem-solving and remote assistance in distributed work contexts. Misalignment produces suboptimal outcomes. Deploying VR for real-time repair introduces safety hazards, while using AR for initial skills training may impose an excessive cognitive load.

C. Comparison with Existing Literature

This review's findings both converge with and extend previous systematic reviews. Wang et al. [82] examined XR in product assembly, identifying similar temporal evolution patterns. However, their assembly focus revealed different technology distributions across workflow phases. In the pre-assembly phase, VR dominated pre-assembly activities, including training, simulation, and planning (30 VR versus 9 AR/MR implementations). Conversely, in the after-assembly phase encompassing maintenance activities, AR/MR showed clear dominance (30 AR/MR versus 8 VR implementations). This pattern aligns with the present review, where AR dominates operational maintenance guidance (57.9% of maintenance studies), highlighting how task characteristics shape technology selection.

Earlier maintenance-specific reviews established a foundational understanding at different maturity stages. Palmarini et al. [83] identified fragmented AR solutions and hardware limitations, concerns that remain partially relevant. Guo et al. [84] focused on VR applications across product lifecycles. Our findings extend these works by providing updated insights on AR hardware evolution, revealing recent VR acceleration post-2020, and documenting the emerging integration of AI and IoT.

The human-centric emphasis in our findings strongly supports the Maintenance 5.0 framework articulated by Psarommatis et al. [85], which expands beyond Industry 4.0's focus on automation and economic goals to encompass human-centricity, sustainability, and resilience. With 44.2% of studies addressing human-centric factors, our findings validate this paradigm shift toward systems that balance technological capability with human well-being, cognitive support, and sustainable operations.

Fang et al. [87] provided a comprehensive analysis of HMD AR in manufacturing, emphasizing the challenges of tracking, registration, and human-computer interaction. Our maintenance-focused analysis confirms these findings and reveals additional barriers, including the complexity of content authoring and the assessment of long-term effectiveness. The convergence among reviews on hardware limitations, registration accuracy, and human factors suggests fundamental rather than transient challenges.

Alam et al. [86] examined XR challenges in Industry 4.0 contexts, focusing on interoperability, adaptability, and cost barriers. Our findings corroborate these, while providing quantitative evidence: only 8.4% of studies included economic evaluation, suggesting systematic underattention to implementation costs in industrial decision-making.

D. Practical Implications

For industry practitioners, this review provides evidence-based guidance on selecting XR technology. Organizations should prioritize AR for procedural guidance and inspection, where demonstrated time reductions of 38% and error reductions of 92.4% provide clear value propositions. VR investments are justified for training applications, particularly for hazardous procedures where 10% efficiency improvements and enhanced knowledge retention offset development costs.

Implementation should follow staged approaches. Initial AR pilots targeting high-complexity maintenance with clear performance metrics enable value demonstration. Critical success factors include simplifying content authoring, robust change management that addresses user acceptance, and evaluation frameworks that capture both quantitative performance and qualitative user experience.

For XR system designers, human-centricity must be an architectural principle. Systems should incorporate adaptive capabilities that adjust to user expertise, task complexity, and environmental conditions. Multi-modal interaction, supporting both gesture and voice commands, accommodates diverse user preferences.

For researchers, this review identifies critical gaps. Longitudinal studies examining sustained XR usage beyond single-session experiments are urgently needed, as are standardized

evaluation protocols enabling cross-study comparisons. Economic models incorporating the total cost of ownership would strengthen industrial adoption decisions. Physiological evaluation methods, currently employed by only 7.4% of studies, offer unique insights complementing traditional performance metrics.

E. Limitations

This review's scope and methodology impose limitations. Including assembly and disassembly tasks broadened the evidence base but may introduce heterogeneity, as assembly emphasizes different demands than traditional maintenance.

Grouping AR and MR reflects common practice but may mask important distinctions in system capabilities. Rapid technological change means findings regarding specific hardware platforms have limited temporal stability—limitations identified for 2015-era HMDs may not apply to current devices.

Most primary studies employed laboratory evaluations with students rather than field deployments with experienced technicians, limiting ecological validity. Small sample sizes and single-session designs restrict understanding of long-term effectiveness and adoption patterns. Publication bias toward positive results may overestimate the effectiveness of XR.

F. Future Research Directions

Immediate research priorities include longitudinal field studies examining XR adoption patterns and sustained usage in authentic industrial settings. Standardized evaluation frameworks would enable systematic comparison across implementations, facilitating meta-analyses. Economic evaluation incorporating comprehensive cost-benefit analyses would strengthen industrial decision-making.

Medium-term priorities should focus on integrating emerging technologies. AI-enhanced XR systems offering context-aware guidance, predictive maintenance support, and adaptive instruction are promising directions aligned with Maintenance 5.0 principles. Digital twin integration connecting XR interfaces with real-time equipment data could enable advanced diagnostic support and predictive capabilities. Multi-modal XR incorporating haptic feedback, spatial audio, and environmental sensing may improve training effectiveness.

Long-term research should examine the ethical implications of XR deployment. Questions regarding worker surveillance, skill deskilling versus augmentation, and equitable access require sustained attention. The development of XR systems that genuinely enhance human capabilities requires critical engagement with human-centered design principles and participatory development approaches.

This systematic literature review demonstrates that XR technologies have progressed beyond proof-of-concept to become viable tools for industrial maintenance. However, significant challenges remain before widespread adoption. Success requires continued attention to human factors, robust evaluation methodologies, and strategic implementation aligned with organizational capabilities and the characteristics of maintenance tasks.

V. Conclusion

This systematic literature review synthesizes evidence from 95 primary studies examining extended reality technologies in industrial maintenance, revealing a rapidly maturing field transitioning from technology-centric implementations toward human-centered systems aligned with Maintenance 5.0 principles. The analysis demonstrates that XR technologies have progressed beyond proof-of-concept to become viable tools that deliver measurable operational improvements while addressing the evolving complexity of modern industrial maintenance.

The research establishes three fundamental contributions. First, it identifies precise technology-task alignments: AR dominates operational guidance and real-time assistance (71.6% of studies), VR excels in training and simulation contexts (38.9%), while MR enables collaborative maintenance scenarios (6.3%). Second, temporal analysis reveals distinct evolutionary phases—from basic visualization (2014-2017), through cognitive enhancement (2018-2021), to AI-integrated adaptive systems (2022-2025)—mirroring the broader industrial transformation toward human-centric maintenance. Third, the prevalence of human-centric factors in 44.2% of studies signals a paradigm shift emphasizing cognitive support, ergonomics, adaptability, and user experience as central design considerations.

This review advances theoretical understanding through two frameworks. The Evaluation Framework expresses XR effectiveness as Evaluation Criteria = f(XR Technology, Task Complexity, Hardware, Application Area, Human-Centric Factors), providing systematic guidance for the design of assessments. The Technology-Task-Context Alignment Model recommends optimal pairings between XR modalities and maintenance requirements, acknowledging that misalignment yields suboptimal outcomes. These frameworks move beyond descriptive analysis to offer prescriptive guidance for implementation decisions.

Practical implications indicate AR investments should target high-complexity procedural guidance, where demonstrated time reductions of 38% and error reductions of 92.4% provide clear value propositions. VR deployment is justified for training applications, particularly those involving hazardous procedures, where the safety benefits and knowledge retention outweigh the development costs. Implementation success requires staged approaches, simplified content authoring, robust change management, and comprehensive evaluation frameworks.

However, these findings must be interpreted in light of persistent limitations. Most primary studies relied on laboratory evaluations with small sample sizes, limiting ecological validity and generalizability. Hardware constraints—including limited field of view, battery life, and tracking accuracy on reflective surfaces—remain unresolved barriers to industrial deployment. Content authoring complexity continues to impose high thresholds for initial adoption, and economic evaluation was addressed by only 8.4% of studies, leaving the cost-effectiveness of XR implementations insufficiently characterized. These observed limitations directly inform the most pressing research priorities. First, longitudinal field studies are needed to address the current reliance on single-session laboratory experiments identified across the reviewed literature. Second, standardized evaluation protocols would

resolve the inconsistencies in assessment methodologies that currently prevent systematic cross-study comparison. Third, comprehensive economic models incorporating total cost of ownership would address the gap in financial analysis documented in this review. Beyond these immediate needs, future research should explore AI-enhanced adaptive systems, digital twin integration, and multi-modal interfaces, while engaging with the ethical implications of sustained XR deployment in industrial workplaces.

Credit Authorship Contribution Statement

Mouad Danane: Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing.

Abdelmajid Elouadi: Methodology, Validation, Supervision.

Youssef Rochdi: Conceptualization, Writing – review & editing.

Declaration of Generative AI and AI Assisted Technologies in the Writing Process

During the preparation of this work, the author(s) used Claude AI (Anthropic) to improve the language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Funding

This research was supported by the CNRST (National Center for Scientific and Technical Research) through the "PhD-Associate Scholarship – PASS" Program.

Declaration of Competing 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 study.

References

[1] Abidi, M. H., Ahmad, A., Darmoul, S., & Al-Ahmari, A. M. (2015). Haptics Assisted Virtual Assembly. *IFAC-PapersOnLine*, 48(3), 100–105. doi:10.1016/j.ifacol.2015.06.065

[2] Alessa, F. M., Alhaag, M. H., Al-harkan, I. M., Ramadan, M. Z., & Alqahtani, F. M. (2023). A Neurophysiological Evaluation of Cognitive Load during Augmented Reality Interactions in Various Industrial Maintenance and Assembly Tasks. *Sensors*, 23(18). doi:10.3390/s23187698

[3] Alessa, F. M., Alhaag, M. H., Al-harkan, I. M., Nasr, M. M., Kaid, H., & Hammami, N. (2024). Evaluating Physical Stress across Task Difficulty Levels in Augmented Reality-Assisted Industrial Maintenance. *Applied Sciences*, 14(1). doi:10.3390/app14010363

[4] Ariansyah, D., Erkoyuncu, J. A., Eimontaite, I., Johnson, T., Oostveen, A.-M., Fletcher, S., & Sharples, S. (2022). A Head Mounted Augmented Reality Design Practice for Maintenance Assembly: Toward Meeting Perceptual and Cognitive Needs of AR Users. *Applied Ergonomics*, 98, 103597. doi:10.1016/j.apergo.2021.103597

[5] Aromaa, S., Aaltonen, I., Kaasinen, E., Elo, J., & Parkkinen, I. (2016). Use of Wearable and Augmented Reality Technologies in Industrial Maintenance Work. In *Proceedings of the International Academic Mindtrek Conference (AcademicMindtrek)* (pp. 235–242). Association for Computing Machinery. doi:10.1145/2994310.2994321

[6] Aromaa, S., Väätänen, A., Hakkarainen, M., & Kaasinen, E. (2018). User Experience and User Acceptance of an Augmented Reality Based Knowledge-Sharing Solution in Industrial Maintenance Work. In C. Falcao & T. Ahram (Eds.), *Advances in Intelligent Systems and Computing* (Vol. 607, pp. 145–156). Springer Verlag. doi:10.1007/978-3-319-60492-3_14

[7] Aschenbrenner, D., Latoschik, M. E., & Schillingz, K. (2016). Industrial Maintenance with Augmented Reality: Two Case Studies. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST)* (pp. 341–342). Association for Computing Machinery. doi:10.1145/2993369.2996305

[8] Bejczy, B., Bozyil, R., Vaičekauskas, E., Krogh Petersen, S. B., Bøgh, S., Hjorth, S. S., & Hansen, E. B. (2020). Mixed Reality Interface for Improving Mobile Manipulator Teleoperation in Contamination Critical Applications. *Procedia Manufacturing*, 51, 620–626. doi:10.1016/j.promfg.2020.10.087

[9] Bernsteiner, R., Ploder, C., Dilger, T., Nigg, J., Spieß, T., & Weichelt, R. (2021). An Analysis of an Augmented Reality Application to Support Service Staff in Industrial Maintenance. In L. Uden, I. Ting, & K. Wang (Eds.), *Communications in Computer and Information Science* (Vol. 1438, pp. 457–467). Springer. doi:10.1007/978-3-030-81635-3_37

[10] Brunzini, A., Grandi, F., Peruzzini, M., & Pellicciari, M. (2021). Virtual Training for Assembly Tasks: A Framework for the Analysis of the Cognitive Impact on Operators. *Procedia Manufacturing*, 55, 527–534. doi:10.1016/j.promfg.2021.10.072

[11] Burova, A., Heinonen, H., Palma, P. B., Keskinen, T., Hakulinen, J., Opas, V., Mäkela, J., Ronkainen, K., Siltanen, S., Raisamo, R., & Turunen, M. (2021). Toward Efficient Academia-Industry Collaboration: A Case Study of Joint VR System Development. In *ACM International Conference Proceeding Series* (pp. 176–185). Association for Computing Machinery. doi:10.1145/3464327.3464367

[12] Burova, A., Mäkela, J., Hakulinen, J., Keskinen, T., Heinonen, H., Siltanen, S., & Turunen, M. (2020). Utilizing VR and Gaze Tracking to Develop AR Solutions for Industrial Maintenance. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. Association for Computing Machinery. doi:10.1145/3313831.3376405

[13] Burova, A., Mäkela, J., Heinonen, H., Palma, P. B., Hakulinen, J., Opas, V., Siltanen, S., Raisamo, R., & Turunen, M. (2022). Asynchronous Industrial Collaboration: How Virtual Reality and Virtual Tools Aid the Process of Maintenance Method Development and Documentation Creation. *Computers in Industry*, 140. doi:10.1016/j.compind.2022.103663

[14] Burova, A., Palma, P. B., Truong, P., Mäkela, J., Heinonen, H., Hakulinen, J., Ronkainen, K., Raisamo, R., Turunen, M., & Siltanen, S. (2022). Distributed Asymmetric Virtual Reality in Industrial Context: Enhancing the Collaboration of Geographically Dispersed Teams in the Pipeline of Maintenance Method Development and Technical Documentation Creation. *Applied Sciences*, 12(8). doi:10.3390/app12083728

[15] Cachada, A., Romero, L., Costa, D., Badikyan, H., Barbosa, J., Leitao, P., Morais, O., Teixeira, C., Azevedo, J., & Moreira, P. M. (2019). Using AR Interfaces to Support Industrial Maintenance Procedures. In *Proceedings of IECON* (pp. 3795–3800). IEEE. doi:10.1109/IECON.2019.8927815

[16] Chang, M. M. L., Ong, S. K., & Nee, A. Y. C. (2017). AR-Guided Product Disassembly for Maintenance and Remanufacturing. *Procedia CIRP*, 61, 299–304. doi:10.1016/j.procir.2016.11.194

[17] Daling, L. M., Tenbrock, M., Isenhardt, I., & Schlittmeier, S. J. (2023). Assemble It like This! – Is AR- or VR-Based Training an Effective Alternative to Video-Based Training in Manual Assembly? *Applied Ergonomics*, 110, 104021. doi:10.1016/j.apergo.2023.104021

[18] De Pace, F., Manuri, F., Sanna, A., & Zappia, D. (2019). Corrigendum: A Comparison Between Two Different Approaches for a Collaborative Mixed-Virtual Environment in Industrial Maintenance. *Frontiers in Robotics and AI*, 6. doi:10.3389/frobt.2019.00051

[19] Elia, V., Gnoni, M. G., & Lanzilotto, A. (2016). Evaluating the Application of Augmented Reality Devices in Manufacturing from a Process Point of View: An AHP Based Model. *Expert Systems with Applications*, 63, 187–197. doi:10.1016/j.eswa.2016.07.006

[20] Erkoyuncu, J., & Khan, S. (2020). Olfactory-Based Augmented Reality Support for Industrial Maintenance. *IEEE Access*, 8, 30306–30321. doi:10.1109/ACCESS.2020.2970220

[21] Erkoyuncu, J. A., del Amo, I. F., Dalle Mura, M., Roy, R., & Dini, G. (2017). Improving Efficiency of Industrial Maintenance with Context Aware Adaptive Authoring in Augmented Reality. *CIRP Annals - Manufacturing Technology*, 66(1), 465–468. doi:10.1016/j.cirp.2017.04.006

[22] Fiorentino, M., Uva, A. E., Gattullo, M., Debernardis, S., & Monno, G. (2014). Augmented Reality on Large Screen for Interactive Maintenance Instructions. *Computers in Industry*, 65(2), 270–278. doi:10.1016/j.compind.2013.11.004

[23] Flotyński, J., Walczak, K., Sobociński, P., & Gałązkiewicz, A. (2021). Knowledge-Based Management of Virtual Training Scenarios. *Procedia Computer Science*, 192, 766–775. doi:10.1016/j.procs.2021.08.079

[24] Freddi, M., & Frizziero, L. (2020). Design for Disassembly and Augmented Reality Applied to a Tailstock. *Actuators*, 9(4), 1–14. doi:10.3390/act9040102

[25] Gattullo, M., Scurati, G. W., Fiorentino, M., Uva, A. E., Ferrise, F., & Bordegoni, M. (2019). Towards Augmented Reality Manuals for Industry 4.0: A Methodology. *Robotics and Computer-Integrated Manufacturing*, 56, 276–286. doi:10.1016/j.rcim.2018.10.001

[26] Gattullo, M., Uva, A. E., Fiorentino, M., Scurati, G. W., & Ferrise, F. (2017). From Paper Manual to AR Manual: Do We Still Need Text? *Procedia Manufacturing*, 11, 1303–1310. doi:10.1016/j.promfg.2017.07.258

[27] Gong, Y., Su, C., & Wei, Y. (2019). An Interactive Task Modeling Method for the Development of Augmented Reality Maintenance Guiding System. In *Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)* (Vol. 1, pp. 136–140). IEEE. doi:10.1109/IHMSC.2019.00039

[28] Gualtieri, L., Revolti, A., & Dallasega, P. (2023). A Human-Centered Conceptual Model for Integrating Augmented Reality and Dynamic Digital Models to Reduce Occupational Risks in Industrial Contexts. *Procedia Computer Science*, 217, 765–773. doi:10.1016/j.procs.2022.12.273

[29] Guler, O., & Yucedag, I. (2018). Developing an CNC Lathe Augmented Reality Application for Industrial Maintanance Training. In *Proceedings of the International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. IEEE. doi:10.1109/ISMSIT.2018.8567255

[30] Hassan, M., Montague, G., Iqbal, M. Z., & Fahey, J. (2024). Virtual Reality-Based Bioreactor Digital Twin for Operator Training. *Digital Chemical Engineering*, 11, 100147. doi:10.1016/j.dche.2024.100147

[31] Havard, V., Baudry, D., Savatier, X., Jeanne, B., Louis, A., & Mazari, B. (2016). Augmented Industrial Maintenance (AIM): A Case Study for Evaluating and Comparing with Paper and Video Media Supports. In L.T. De Paolis & A. Mongelli (Eds.), *Lecture Notes in Computer Science* (Vol. 9768, pp. 302–320). Springer Verlag. doi:10.1007/978-3-319-40621-3_22

[32] Heinonen, H., Burova, A., Siltanen, S., Lähteenmäki, J., Hakulinen, J., & Turunen, M. (2022). Evaluating the Benefits of Collaborative VR Review for Maintenance Documentation and Risk Assessment. *Applied Sciences*, 12(14). doi:10.3390/app12147155

[33] Heinz, M., Dhiman, H., & Röcker, C. (2018). A Multi-Device Assistive System for Industrial Maintenance Operations. In *Lecture Notes in Computer Science* (Vol. 11015, pp. 239–247). Springer Verlag. doi:10.1007/978-3-319-99740-7_16

[34] Isik, B., Isik, G. E., & Zilka, M. (2024). Integrating Extended Reality in Industrial Maintenance: A Game-Based Framework for Compressed Air System Training. In *Procedia Computer Science* (Vol. 232, pp. 483–492). Elsevier B.V. doi:10.1016/j.procs.2024.01.048

[35] Kalkan, Ö. K., Karabulut, Ş., & Höke, G. (2021). Effect of Virtual Reality-Based Training on Complex Industrial Assembly Task Performance. *Arabian Journal for Science and Engineering*, 46(12), 12697–12708. doi:10.1007/s13369-021-06138-w

[36] Laviola, E., Gattullo, M., Evangelista, A., Fiorentino, M., & Uva, A. E. (2023). In-Situ or Side-by-Side? A User Study on Augmented Reality Maintenance Instructions in Blind Areas. *Computers in Industry*, 144, 103795. doi:10.1016/j.compind.2022.103795

[37] Leutert, F., & Schilling, K. (2018). Projector-Based Augmented Reality for Telemaintenance Support. *IFAC-PapersOnLine*, 51(11), 502–507. doi:10.1016/j.ifacol.2018.08.368

[38] Liu, C., Zhu, H., Tang, D., Nie, Q., Zhou, T., Wang, L., & Song, Y. (2022). Probing an Intelligent Predictive Maintenance Approach with Deep Learning and Augmented Reality for Machine Tools in IoT-Enabled Manufacturing. *Robotics and Computer-Integrated Manufacturing*, 77, 102357. doi:10.1016/j.rcim.2022.102357

[39] Loch, F., Ziegler, U., & Vogel-Heuser, B. (2018). Integrating Haptic Interaction into a Virtual Training System for Manual Procedures in Industrial Environments. *IFAC-PapersOnLine*, 51(11), 60–65. doi:10.1016/j.ifacol.2018.08.235

[40] Loch, F., Ziegler, U., & Vogel-Heuser, B. (2019). Using Real-Time Feedback in a Training System for Manual Procedures. *IFAC-PapersOnLine*, 52(19), 241–246. doi:10.1016/j.ifacol.2019.12.089

[41] Mei, Y., Liu, Y., Sun, C., Wang, X., Wang, D., Yuan, L., & Tan, J. (2023). Multi-Stage Rotors Assembly of Turbine-Based Combined Cycle Engine Based on Augmented Reality. *Advanced Engineering Informatics*, 58, 102160. doi:10.1016/j.aei.2023.102160

[42] Moghaddam, M., Wilson, N. C., Modestino, A. S., Jona, K., & Marsella, S. C. (2021). Exploring Augmented Reality for Worker Assistance versus Training. *Advanced Engineering Informatics*, 50, 101410. doi:10.1016/j.aei.2021.101410

[43] Mompeu, G., Danglade, F., Mérienne, F., & Guillet, C. (2024). Methodology for Augmented Reality-Based Adaptive Assistance in Industry. *Computers in Industry*, 154, 104021. doi:10.1016/j.compind.2023.104021

[44] Monetti, F. M., de Giorgio, A., Yu, H., Maffei, A., & Romero, M. (2022). An Experimental Study of the Impact of Virtual Reality Training on Manufacturing Operators on Industrial Robotic Tasks. *Procedia CIRP*, 106, 33–38. doi:10.1016/j.procir.2022.02.151

[45] Mourtzis, D., Angelopoulos, J., & Boli, N. (2018). Maintenance Assistance Application of Engineering to Order Manufacturing Equipment: A Product Service System (PSS) Approach. *IFAC-PapersOnLine*, 51(11), 217–222. doi:10.1016/j.ifacol.2018.08.263

[46] Mourtzis, D., Xanthi, F., & Zogopoulos, V. (2019). An Adaptive Framework for Augmented Reality Instructions Considering Workforce Skill. *Procedia CIRP*, 81, 363–368. doi:10.1016/j.procir.2019.03.063

[47] Mura, M. D., & Dini, G. (2021). A Proposal of an Assembly Workstation for Car Panel Fitting Aided by an Augmented Reality Device. *Procedia CIRP*, 103, 225–230. doi:10.1016/j.procir.2021.10.036

[48] Neges, M., Adwernat, S., & Abramovici, M. (2018). Augmented Virtuality for Maintenance Training Simulation under Various Stress Conditions. *Procedia Manufacturing*, 19, 171–178. doi:10.1016/j.promfg.2018.01.024

[49] Numfu, M., Riel, A., & Noël, F. (2020). A Concept for Virtual Reality Based Industrial Maintenance Training Preparation. In *Communications in Computer and Information Science* (Vol. 1251, pp. 820–829). Springer. doi:10.1007/978-3-030-56441-4_62

[50] Numfu, M., Riel, A., & Noel, F. (2019). Virtual Reality Based Digital Chain for Maintenance Training. *Procedia CIRP*, 84, 1069–1074. doi:10.1016/j.procir.2019.04.268

[51] Ortega, M., Ivorra, E., Juan, A., Venegas, P., Martínez, J., & Alcañiz, M. (2021). Mantra: An Effective System Based on Augmented Reality and Infrared Thermography for Industrial Maintenance. *Applied Sciences*, 11(1), 1–26. doi:10.3390/app11010385

[52] Osti, F., Ceruti, A., Liverani, A., & Caligiana, G. (2017). Semi-Automatic Design for Disassembly Strategy Planning: An Augmented Reality Approach. *Procedia Manufacturing*, 11, 1481–1488. doi:10.1016/j.promfg.2017.07.279

[53] Oulefki, A., Himeur, Y., Trongtirakul, T., Amara, K., Agaian, S., Benbelkacem, S., Guerroudji, M. A., Zemmouri, M., Ferhat, S., Zenati, N., Atalla, S., & Mansoor, W. (2024). Detection and Analysis of Deteriorated Areas in Solar PV Modules Using Unsupervised Sensing Algorithms and 3D Augmented Reality. *Heliyon*, 10(6), e27973. doi:10.1016/j.heliyon.2024.e27973

[54] Park, K.-B., Kim, M., Choi, S. H., & Lee, J. Y. (2020). Deep Learning-Based Smart Task Assistance in Wearable Augmented Reality. *Robotics and Computer-Integrated Manufacturing*, 63, 101887. doi:10.1016/j.rcim.2019.101887

[55] Quandt, M., Beinke, T., & Freitag, M. (2020). User-Centered Evaluation of an Augmented Reality-Based Assistance System for Maintenance. *Procedia CIRP*, 93, 921–926. doi:10.1016/j.procir.2020.03.053

[56] Richard, K., Havard, V., His, J., & Baudry, D. (2021). INTERVALES: INTERactive Virtual and Augmented Framework for industriaL Environment and Scenarios. *Advanced Engineering Informatics*, 50, 101425. doi:10.1016/j.aei.2021.101425

[57] Ritucci, A. L., Frizziero, L., & Liverani, A. (2021). Maintainability Approach: Hydraulic Pump with External Gears Explored with Design for Disassembly and Augmented Reality. *Applied Sciences*, 11(2), 1–23. doi:10.3390/app11020666

[58] Sanna, A., Manuri, F., Piumatti, G., Paravati, G., Lamberti, F., & Pezzolla, P. (2015). A Flexible AR-Based Training System for Industrial Maintenance. In *Lecture Notes in Computer Science* (Vol. 9254, pp. 314–331). Springer Verlag. doi:10.1007/978-3-319-22888-4_23

[59] Sattarpanah Karganroudi, S., Silva, R. E., Chahdi El Ouazani, Y., Aminzadeh, A., Dimitrova, M., & Ibrahim, H. (2022). A Novel Assembly Process Guidance Using Augmented Reality for a Standalone Hybrid Energy System. *The International Journal of Advanced Manufacturing Technology*, 122(7–8), 3425–3445. doi:10.1007/s00170-022-10122-5

[60] Schlagowski, R., Merkel, L., & Meitinger, C. (2017). Design of an Assistant System for Industrial Maintenance Tasks and Implementation of a Prototype Using Augmented Reality. In *Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 294–298). IEEE. doi:10.1109/IEEM.2017.8289899

[61] Scurati, G. W., Gattullo, M., Fiorentino, M., Ferrise, F., Bordegoni, M., & Uva, A. E. (2018). Converting Maintenance Actions into Standard Symbols for Augmented Reality Applications in Industry 4.0. *Computers in Industry*, 98, 68–79. doi:10.1016/j.compind.2018.02.001

[62] Serras, M., Garcia-Sardina, L., Simoes, B., Alvarez, H., & Arambarri, J. (2020). AREVA: Augmented Reality Voice Assistant for Industrial Maintenance. *Procesamiento del Lenguaje Natural*, 65, 135–138. doi:10.26342/2020-65-21

[63] Shamsuzzoha, A., Helo, P., Kankaanpää, T., Toshev, R., & Tuan, V. V. (2018). Applications of Virtual Reality in Industrial Repair and Maintenance. In *Proceedings of the International Conference on Industrial Engineering and Operations Management* (pp. 1890–1898). IEOM Society.

[64] Shamsuzzoha, A., Toshev, R., Vu Tuan, V., Kankaanpaa, T., & Helo, P. (2021). Digital Factory–Virtual Reality Environments for Industrial Training and Maintenance. *Interactive Learning Environments*, 29(8), 1339–1362. doi:10.1080/10494820.2019.1628072

[65] Siew, C. Y., Nee, A. Y. C., & Ong, S. K. (2019). Improving Maintenance Efficiency with an Adaptive AR-Assisted Maintenance System. In *ACM International Conference Proceeding Series* (pp. 74–78). Association for Computing Machinery. doi:10.1145/3351180.3351203

[66] Talami, F., de Barros, A., Silva, P., Vasconcelos, A., & Santos, G. (2023). AR in Remote Maintenance: Empirical User Research with Dyads. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (pp. 425–431). IEEE. doi:10.1109/VRW58643.2023.00092

[67] Tripicchio, P., Avizzano, C. A., & Bergamasco, M. (2019). A 6-DOF Haptic Manipulation System to Verify Assembly Procedures on CAD Models. *Procedia Manufacturing*, 38, 1292–1299. doi:10.1016/j.promfg.2020.01.161

[68] Wang, Z., Zhang, X., Li, L., Zhou, Y., Lu, Z., Dai, Y., Liu, C., Su, Z., Bai, X., & Billinghurst, M. (2023). Evaluating Visual Encoding Quality of a Mixed Reality User Interface for Human–Machine Co-Assembly in Complex Operational Terrain. *Advanced Engineering Informatics*, 58, 102171. doi:10.1016/j.aei.2023.102171

[69] Weltin, M., Lucke, D., & Jooste, J. L. (2023). Automatic Content Creation System for Augmented Reality Maintenance Applications for Legacy Machines. *Procedia CIRP*, 120, 750–755. doi:10.1016/j.procir.2023.09.070

[70] Zhang, X., Bai, X., Zhang, S., He, W., Wang, S., Yan, Y., & Wang, P., Liu, L. (2024). A Novel Mixed Reality Remote Collaboration System with Adaptive Generation of Instructions. *Computers & Industrial Engineering*, 110353. doi:10.1016/j.cie.2024.110353

[71] Zhang, X., Bai, X., Zhang, S., He, W., Wang, S., Yan, Y., Yu, Q., & Liu, L. (2023). A Novel MR Remote Collaboration System Using 3D Spatial Area Cue and Visual Notification. *Journal of Manufacturing Systems*, 67, 389–409. doi:10.1016/j.jmsy.2023.02.013

[72] Zheng, L., Liu, X., An, Z., Li, S., & Zhang, R. (2020). A Smart Assistance System for Cable Assembly by Combining Wearable Augmented Reality with Portable Visual Inspection. *Virtual Reality & Intelligent Hardware*, 2(1), 12–27. doi:10.1016/j.vrih.2019.12.002

[73] Velosa, J. D., Cobo, L., Castillo, F., & Castillo, C. (2018). Methodological Proposal for Use of Virtual Reality VR and Augmented Reality AR in the Formation of Professional Skills in Industrial Maintenance and Industrial Safety. In *Lecture Notes in Networks and Systems* (Vol. 22, pp. 987–1000). Springer. doi:10.1007/978-3-319-64352-6_92

[74] Nagy, A., Spyridis, Y., Mills, G. J., & Argyriou, V. (2024). User Experience Evaluation of AR Assisted Industrial Maintenance and Support Applications. In *Proceedings of the International Conference on Virtual Reality (ICVR)* (pp. 270–277). doi:10.1109/ICVR62393.2024.10869099

[75] Stover, D., & Bowman, D. (2024). TAGGAR: General-Purpose Task Guidance from Natural Language in Augmented Reality Using Vision-Language Models. doi:10.1145/3677386.3682095

[76] Mompeu, G., Danglade, F., Mérienne, F., & Guillet, C. (2024). Methodology for Augmented Reality-Based Adaptive Assistance in Industry. *Computers in Industry*, 154, 104021. doi:10.1016/j.compind.2023.104021

[77] Obermair, F., Althaler, J., Wolfartsberger, J., Niedermayr, D., & Feichtenschlager, H.-P. (2024). Assembly Training for Electric Motor Wiring with Virtual Reality Compared to Paper Instructions: Experiment and Analysis. *Procedia CIRP*, 130, 398–406. doi:10.1016/j.procir.2024.10.106

[78] Padovano, A., Cardamone, M., & Klaess, J. (2024). Empowering Operator 5.0: Human-Centric Design of an Augmented Reality Tool for a Learning Factory. In *6th IFAC Workshop on Advanced Maintenance Engineering, Services and Technology (AMEST)* (Vol. 58, pp. 180–185). doi:10.1016/j.ifacol.2024.08.117

[79] Costa, A., Miranda, J., Dias, D., Dinis, N., Romero, L., Faria, P. M., Smart Maintenance Solutions: AR- and VR-Enhanced Digital Twin Powered by FIWARE, *Sensors*, 25(3), 2025. doi:10.3390/s25030845.

[80] Işık, B., Işık, G. E., Zilka, M., Game-Based Learning for Industrial Maintenance: A Unity 3D Educational Game of Compressed Air System Training, *Procedia Computer Science*, 253, 784–793, 2025. doi:10.1016/j.procs.2025.01.140.

[81] Dianatfar, M., Järvenpää, E., Siltala, N., Lanz, M., Template Concept for VR Environments: A Case Study in VR-Based Safety Training for Human–Robot Collaboration, *Robotics and Computer-Integrated Manufacturing*, 94, 102973, 2025. doi:10.1016/j.rcim.2025.102973.

[82] Wang, B., Zheng, L., Wang, Y., Fang, W., Wang, L., Towards the industry 5.0 frontier: Review and prospect of XR in product assembly, *Journal of Manufacturing Systems*, 74, 777–811, 2024. doi:10.1016/j.jmsy.2024.05.002.

[83] Palmarini, R., Erkoyuncu, J. A., Roy, R., Torabmostaedi, H., A systematic review of augmented reality applications in maintenance, *Robotics and Computer-Integrated Manufacturing*, 49, 215–228, 2018. doi:10.1016/j.rcim.2017.06.002.

[84] Guo, Z., Zhou, D., Zhou, Q., Zhang, X., Geng, J., Zeng, S., Lv, C., Hao, A., Applications of virtual reality in maintenance during the industrial product lifecycle: A systematic review, *Journal of Manufacturing Systems*, 56, 525–538, 2020. doi:10.1016/j.jmsy.2020.07.007.

[85] Psarommatis, F., May, G., Azamfirei, V., Envisioning maintenance 5.0: Insights from a systematic literature review of Industry 4.0 and a proposed framework, *Journal of Manufacturing Systems*, 68, 376–399, 2023. doi:10.1016/j.jmsy.2023.04.009.

[86] Alam, N., Saha, N., Gadow, V., Harik, R., Ryu, J., Role of extended reality (XR) technologies in maintenance operations: Trends, challenges, and integration in industry 4.0, *Manufacturing Letters*, 44, 1545–1557, 2025. doi:10.1016/j.mfglet.2025.06.174.

[87] Fang, W., Chen, L., Zhang, T., Chen, C., Teng, Z., Wang, L., Head-mounted display augmented reality in manufacturing: A systematic review, *Robotics and Computer-Integrated Manufacturing*, 83, 102567, 2023. doi:10.1016/j.rcim.2023.102567.

[88] Esmaeilian, B., Behdad, S., Wang, B., The evolution and future of manufacturing: A review, *Journal of Manufacturing Systems*, 39, 79–100, 2016.

[89] Salawu, E. Y., Awoyemi, O. O., Akerekan, O. E., Afolalu, S. A., Kayode, J. F., Ongbali, S. O., Airewa, I., Edun, B. M., Impact of maintenance on machine reliability: A review, In *E3S Web of Conferences*, 430, 01226, 2023. EDP Sciences.

[90] Achouch, M., Dimitrova, M., Ziane, K., Sattarpanah Karganroudi, S., Dhouib, R., Ibrahim, H., Adda, M., On predictive maintenance in industry 4.0: Overview, models, and challenges, *Applied Sciences*, 12(16), 8081, 2022. MDPI.

[91] Zonta, T., Da Costa, C. A., da Rosa Righi, R., de Lima, M. J., Da Trindade, E. S., Li, G. P., Predictive maintenance in the Industry 4.0: A systematic literature review, *Computers & Industrial Engineering*, 150, 106889, 2020. Elsevier.

[92] Nahavandi, S., Industry 5.0—A human-centric solution, *Sustainability*, 11(16), 4371, 2019. MDPI.

[93] Casini, M., Extended reality for smart building operation and maintenance: A review, *Energies*, 15(10), 3785, 2022. MDPI.

[94] Coupry, C., Noblecourt, S., Richard, P., Baudry, D., Bigaud, D., BIM-Based digital twin and XR devices to improve maintenance procedures in smart buildings: A literature review, *Applied Sciences*, 11(15), 6810, 2021. MDPI.

[95] Yazdi, M., Augmented reality (AR) and virtual reality (VR) in maintenance training, In *Advances in Computational Mathematics for Industrial System Reliability and Maintainability*, 169–183, 2024. Springer.

Design and Implementation of a Trust-Aware Fog Computing Simulation Framework Using WorkflowSim

Chia Chuan Wu[1], Selvakumar Manickam[2]*, Shams Ul Arfeen Laghari[3], Shankar Karuppayah[4]
Cybersecurity Research Centre (CYRES), Universiti Sains Malaysia,
George Town, Penang, 11800, Malaysia[1,2,4]
Faculty of Vocational Studies, Universitas Brawijaya, Kota Malang, Indonesia[2]
School of ICT, Faculty of Engineering Design Information and Communications Technology (EDICT),
Bahrain Polytechnic, Isa Town, Bahrain[3]

***Abstract*—Fog computing extends cloud capabilities to the network edge, providing low-latency services for mission-critical applications such as healthcare and industrial automation. However, validating security-aware scheduling policies in this distributed paradigm remains a challenge due to the lack of native support for trust modeling in standard simulation tools like CloudSim and iFogSim. Existing simulators focus primarily on resource provisioning, cost, and energy metrics, often neglecting "Trust" and "Data Confidentiality" as first-class simulation parameters. This study presents the design and implementation of a trust-aware extension for WorkflowSim. We detail the software architecture modifications required to support attribute-based trust verification, secure task fragmentation, and encryption overhead modeling. Unlike standard simulators, our extended framework treats Trust as a dynamic entity in the Virtual Machine (VM) and Task class hierarchy. We validate the framework's correctness through unit tests and scenario-based trace analysis, demonstrating its ability to accurately model security-performance trade-offs in heterogeneous fog environments. Experimental results indicate that enabling trust-aware scheduling increases makespan by approximately 19% and cost by 16%, while achieving a security score of 0.88 compared to 0.65 for the baseline. The simulation engine overhead remains below 7.1% even for 1000-task workflows. The source code and design patterns presented provide researchers with a robust, extensible tool for evaluating secure scheduling algorithms without the cost of physical testbeds.**

***Keywords*—Fog computing; simulation tools; WorkflowSim; trust modeling; task scheduling; software architecture**

I. Introduction

The rapid proliferation of the Internet of Things (IoT) [1], [2] has generated unprecedented volumes of data at the network edge. Traditional cloud computing architectures [3], while offering vast storage and resources, often fail to meet the stringent latency requirements of modern applications [4], [5]. Fog computing has emerged as a decentralized solution, bridging the gap between IoT devices and the cloud [6], [7]. However, the heterogeneous and distributed nature of fog nodes introduces significant security vulnerabilities [8], particularly regarding data confidentiality and node trustworthiness.

This raises a fundamental research question: How can a standard workflow simulation framework be extended to natively support dynamic trust evaluation and data confidentiality modeling for secure fog computing research?

Designing secure scheduling algorithms for fog environments requires rigorous testing. However, deploying physical testbeds with hundreds of fog nodes is often cost-prohibitive and lacks reproducibility. Consequently, researchers rely heavily on simulation tools. While established simulators like CloudSim [14] and iFogSim [17] provide robust models for energy consumption and network latency, they notably lack native support for security-aware modeling. They treat all computing nodes as equally trustworthy, which is an unrealistic assumption in open fog environments where nodes may be malicious or compromised.

A. Motivation: Security Challenges in Fog Computing

In traditional cloud paradigms, data is processed in centralized, physically secure data centers managed by trusted providers (e.g., AWS, Azure). In contrast, fog computing relies on a collaborative model [9], [10] where computation is offloaded to edge devices (e.g., routers, gateways, base stations) often owned by third parties or deployed in public spaces. This architectural shift introduces several critical security challenges:

1) Lack of physical control: Edge nodes are susceptible to physical tampering and side-channel attacks [11], making it risky to process sensitive data (e.g., patient health records) on them without verification.

2) Dynamic trustworthiness: A node that is trustworthy at time t may become compromised at time $t+\Delta t$. Static security policies are insufficient; the system requires continuous trust evaluation [12].

3) Regulatory compliance: Legal frameworks such as GDPR and HIPAA [13] impose strict requirements on data residency and processing locations. Simulators must effectively model these geospatial and trust constraints to be useful for compliance auditing.

Existing simulation tools fail to capture these nuances, typically treating security as a binary state (encrypted/unencrypted) rather than a dynamic attribute influenced by node behavior and environmental factors [7], [8]. This limitation hinders the development of adaptive security algorithms.

*Corresponding author

B. Research Contributions

To address this gap, this study presents the design and implementation of a security-oriented extension to **WorkflowSim** [16]. WorkflowSim was chosen as the base platform due to its support for complex scientific workflow dependencies (DAGs), which are common in data-intensive fog applications.

The primary contributions of this study are:

1) Architectural extension: We propose a modified class hierarchy for WorkflowSim that integrates "Trust" and "Confidentiality" as core attributes of tasks and virtual machines.

2) Trust manager module: We implement a dedicated `TrustManager` component that decouples security logic from the core scheduling engine, enabling dynamic trust updates based on transaction success/failure.

3) Validation of tool correctness: We demonstrate the simulator's capability to enforce security policies (e.g., blocking sensitive tasks from low-trust nodes) through comprehensive trace analysis, rather than just algorithmic performance.

The remainder of this study is organized as follows: Section II reviews existing simulation tools. Section III presents the system model and mathematical formulation. Section IV details the proposed object-oriented system architecture. Section V provides the Java implementation details of the key components. Section VI presents the validation results of the simulation framework, and Section VII provides the discussion. Section VIII concludes the work.

II. Related Work on Fog Simulation

Simulation is a cornerstone of fog computing research due to the complexity and cost of physical testbeds. Several tools have been developed to model different aspects of the infrastructure [18], [19]. This section reviews widely used simulators and identifies the gap that necessitates our trust-aware extension.

A. CloudSim and its Derivatives

CloudSim [14], [15] is the most widely cited toolkit for cloud environments. It provides basic classes for data centers, hosts, and VMs. However, its scheduling logic is predominantly designed for independent tasks (Bag-of-Tasks) and centralized resource brokering. It lacks the concept of "workflow dependencies" (Parent-Child tasks) and decentralized trust management required for fog scenarios.

B. WorkflowSim

WorkflowSim [16] extends CloudSim by adding a Workflow Engine that parses Directed Acyclic Graphs (DAGs). It introduces a `ClusteringEngine` to merge small tasks, reducing simulation overhead. While WorkflowSim excels at modeling task dependencies, its standard distribution does not include security parameters. Tasks are scheduled based solely on completion time or cost, ignoring the security posture of the target node.

C. iFogSim

iFogSim [17] was specifically designed for modeling hierarchical fog and edge computing topologies. It excels at simulating data processing pipelines (sensors to actuators) and energy consumption. However, iFogSim's resource management policies are hard-coded into the `FogDevice` class. Implementing dynamic trust models often requires invasive modification of the core engine, which hampers upgradability and modularity.

D. EdgeCloudSim and PureEdgeSim

EdgeCloudSim [20] extends CloudSim to cover mobility support and network link modeling, crucial for mobile edge computing (MEC) scenarios. Similarly, PureEdgeSim [21] allows for simulation of peer-to-peer edge environments. While both tools provide excellent network fidelity, they lack specific constructs for *Data Confidentiality* and *Trust-Based Scheduling*. They focus primarily on task offloading decisions based on signal strength or battery levels, rather than security attributes.

E. FogNetSim++

Unlike the Java-based tools mentioned above, FogNetSim++ [22] is built on OMNeT++, offering highly detailed network protocol simulation (TCP/IP level). While arguably superior for network analysis, it lacks the high-level workflow scheduling abstractions needed for application-layer research. Furthermore, its C++ codebase is less accessible to the broad community of researchers familiar with the CloudSim Java ecosystem.

F. Trust Models and Reputation Systems

Beyond simulators, numerous trust models have been proposed for distributed systems. The Beta Reputation System (BRS) [23] is a probabilistic trust model based on the beta probability density function, widely used in peer-to-peer (P2P) networks. It calculates reputation by aggregating positive and negative interactions. EigenTrust [24] is another prominent algorithm designed to calculate the global trust value of each peer in a network by computing the left principal eigenvector of a matrix of normalized local trust values.

In the context of fog computing, recent works [25]–[27] focus on lightweight, context-aware trust management. However, these models are typically evaluated using custom, non-reproducible Python scripts or MATLAB simulations, rather than integrated into standard, community-accepted simulation frameworks like WorkflowSim. This disconnect limits the ability of the broader research community to validate, compare, and extend these trust algorithms.

In our proposed framework, we adopt the Beta Reputation System over EigenTrust for several reasons. First, EigenTrust requires iterative convergence across a fully connected peer network to compute global trust, which does not map naturally to the hierarchical fog topology (Mist-Fog-Cloud) in our model. Second, the Beta model operates in $O(1)$ per update, making it suitable for real-time scheduling decisions. Third, the Beta model directly accommodates the binary interaction outcomes (task success/failure) central to our trust formulation.

Importantly, our `TrustManager` interface is designed to be modular—researchers can substitute EigenTrust or blockchain-based trust models [28] by implementing the same interface without modifying the scheduler code.

G. Comparative Analysis

Table I summarizes the capabilities of existing simulators compared to the proposed WorkflowSim extension. While tools like iFogSim and EdgeCloudSim offer robust network and energy modeling, they lack native support for dynamic trust management and data confidentiality attributes in their core task scheduling entities.

H. The Need for a Trust-Aware Extension

Current research into "Secure Fog Scheduling" often relies on ad-hoc, one-off modifications to these simulators. Most researchers hard-code security checks directly into their scheduling algorithms. This study proposes a structural, architectural extension to WorkflowSim that decouples trust logic from scheduling logic, providing a reusable framework for future security research. Use of WorkflowSim ensures support for complex DAG workflows, which is a key requirement for modern data-intensive fog applications [31].

III. System Model and Mathematical Formulation

This section formalizes the system entities, task characteristics, and the multi-objective optimization problem addressed by our proposed framework. The mathematical rigor introduced here ensures that the simulation results are grounded in verifiable theoretical models.

A. Fog Computing Environment

We model the fog computing environment as a hierarchical graph $G_{fog} = (V, E)$, where V represents the set of computing nodes and E represents the communication links. The set of nodes V is partitioned into three tiers:

$$V = V_{cloud} \cup V_{fog} \cup V_{mist} \tag{1}$$

Each node $v_j \in V$ is characterized by a tuple:

$$v_j = \langle P_j, M_j, C_j, T_j, E_j \rangle \tag{2}$$

where,

- P_j: Processing capacity in Million Instructions Per Second (MIPS).
- M_j: Available memory (RAM).
- C_j: Cost per unit time of usage.
- T_j: Current trust score, where $0 \leq T_j \leq 1$.
- E_j: Energy consumption coefficient (Joules/instruction).

B. Workflow Application Model

A scientific workflow W is represented as a Directed Acyclic Graph (DAG) [29] $W = (\mathcal{T}, \mathcal{D})$, where $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ is the set of tasks and $\mathcal{D}$ represents the data dependencies between them. An edge $(t_i, t_k) \in \mathcal{D}$ implies that task t_k cannot start execution until t_i has completed and transferred its output data.

Each task $t_i \in \mathcal{T}$ is defined as:

$$t_i = \langle L_i, D_{in,i}, D_{out,i}, S_i, \tau_i \rangle \tag{3}$$

where,

- L_i: Task length in Million Instructions (MI).
- $D_{in,i}, D_{out,i}$: Input and output data sizes.
- S_i: Security sensitivity level (1: Public, 2: Internal, 3: Confidential).
- τ_i: Maximum tolerable latency (deadline).

C. Trust and Security Formulation

The trustworthiness of a fog node is not static but evolves based on its historical behavior. We employ a Beta Probability Density Function to model trust, as it naturally represents binary outcomes (successful vs. failed interactions).

The trust score T_j of node v_j is updated as:

$$T_j = \frac{\alpha_j + 1}{\alpha_j + \beta_j + 2} \tag{4}$$

where, α_j is the accumulated number of successful task executions (cooperative behavior) and β_j is the number of failed or malicious interactions.

To ensure that recent behavior is weighted more heavily, we apply a time-decay factor λ ($0 < \lambda < 1$):

$$\alpha_j(t) = \alpha_j(t-1) \cdot \lambda + \delta_{success} \tag{5}$$

$$\beta_j(t) = \beta_j(t-1) \cdot \lambda + \delta_{failure} \tag{6}$$

A task t_i with sensitivity S_i can only be scheduled on node v_j if the node satisfies the security constraint:

$$T_j \geq T_{req}(S_i) \tag{7}$$

where, $T_{req}(S_i)$ is a mapping function defining the minimum trust threshold for each sensitivity level (e.g., Level 3 requires $T_j \geq 0.9$).

TABLE I. COMPARISON OF FOG COMPUTING SIMULATORS

Feature	CloudSim	iFogSim	EdgeCloudSim	Proposed
Base Platform	Java	CloudSim	CloudSim	WorkflowSim
Workflow DAGs	No	Limited	No	**Yes**
Trust Model	No	No	No	**Yes**
Confidentiality	No	No	No	**Yes**
Fog Hierarchy	No	Yes	Yes	**Yes**
Network Model	Simple	Detailed	Detailed	**Simple**

D. Encryption Overhead Modeling

Unlike standard simulations that assume security is instantaneous, our model accounts for the computational overhead of encryption. The execution time $ET_{i,j}$ of task t_i on node v_j is composed of processing time and cryptographic overhead:

$$ET_{i,j} = \frac{L_i}{P_j} + \mathbb{I}(S_i > 1) \cdot \frac{D_{in,i} + D_{out,i}}{R_{enc,j}} \tag{8}$$

where, $R_{enc,j}$ is the encryption throughput of node v_j, and $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if the task requires encryption (Sensitivity > 1), and 0 otherwise.

E. Objective Function

The scheduling problem addresses a multi-objective optimization goal [30]: minimizing the makespan (completion time) and total cost while satisfying security constraints. The objective function F is defined as:

$$\text{minimize } F = w_1 \cdot \frac{M_{makespan}}{M_{max}} + w_2 \cdot \frac{C_{total}}{C_{max}} + w_3 \cdot \frac{1}{S_{avg}} \tag{9}$$

subject to:

$$\forall t_i \in \mathcal{T}, \quad T_{assigned_node} \geq T_{req}(S_i) \tag{10}$$

where, w_1, w_2, w_3 are user-defined weights prioritizing time, cost, and security respectively. In the experiments reported in Section VI, we use $w_1 = 0.4$, $w_2 = 0.3$, and $w_3 = 0.3$, reflecting a slight priority on makespan while maintaining balanced consideration of cost and security. In practice, safety-critical applications (e.g., healthcare) should increase w_3, while cost-sensitive batch workloads should increase w_2.

IV. PROPOSED SYSTEM ARCHITECTURE

The proposed simulation framework involves a significant extension of the standard WorkflowSim class hierarchy. The design follows the Object-Oriented Programming (OOP) principles of inheritance and encapsulation to ensure compatibility with the underlying CloudSim engine.

A. Design Philosophy: Modularity and Extensibility

A key design principle of this framework is the decoupling of *Security Logic* from *Scheduling Logic*. In many existing ad-hoc implementations, trust calculations are embedded directly within the scheduler's `run()` loop. This creates tight coupling, making it difficult to swap trust models (e.g., changing from a Beta Reputation System to an EigenTrust model).

Our architecture addresses this by treating the `TrustManager` as a standalone component with a defined interface. The scheduler operates as a client of this manager.

- Encapsulation: The `TrustManager` hides the complexity of how trust is calculated (e.g., history windows, decay factors).
- Interchangeability: Researchers can implement a new `BlockchainTrustManager` that extends the base class without modifying a single line of the `FATSScheduler` code.
- Composition over Inheritance: We utilize object composition where the Scheduler *has-a* TrustManager, rather than forcing a specific inheritance hierarchy.

B. Supported Infrastructure Topologies

The proposed extension supports a three-tier hierarchical architecture, which is critical for realistic fog computing simulations. By modifying the `Datacenter` configuration parser, the tool allows users to define heterogeneous nodes with distinct computational and security profiles:

- Mist Tier: Represents edge devices such as gateways and smart controllers. These nodes are characterized by low computational power, but high physical proximity to data sources. In our model, they are initialized with high base trust due to direct ownership.
- Fog Tier: Represents regional processing units like cloudlets or base stations. These nodes offer moderate computational capacity and act as intermediaries. Their trust levels are dynamic and can degrade over time.
- Cloud Tier: Represents centralized data centers with virtually unlimited resources but high latency. While computationally powerful, they are treated as "semi-trusted" for sensitive data due to third-party management.

The simulation framework allows researchers to configure bandwidth and latency constraints between these tiers, as detailed in Table II.

TABLE II. DEFAULT NETWORK CONFIGURATION PARAMETERS

Link Type	Latency (ms)	Bandwidth (Mbps)
IoT ↔ Mist	2–5	10
Mist ↔ Fog	5–15	50
Fog ↔ Cloud	50–100	100

C. Extended Class Hierarchy

Fig. 1 (see description below) illustrates the core classes modified in this framework.

1) SecureJob (extends job): The standard `Job` class in WorkflowSim represents a computational task. We extend this to `SecureJob` to include sensitivity attributes. This allows researchers to model diverse security requirements without modifying the core engine.

- `confidentialityLevel` (int): A generic integer (e.g., 1-3) representing data sensitivity.
- `encryptionOverhead` (double): The estimated CPU time required for crypto operations.
- `requiredTrust` (double): The minimum trust threshold for this task type.

2) TrustedVM (Extends CondorVM): The `CondorVM` class represents a virtual machine. We extend this to `TrustedVM` to add:

- `trustScore` (double): A dynamic value representing node reliability.
- `securityTier` (enum): Categorization of nodes (e.g., MIST/FOG/CLOUD).
- `encryptionSupport` (boolean): Flag for hardware crypto acceleration.

3) TrustManager (Interface and Implementation): To decouple security logic, we introduce a `TrustManager` component. This standalone module acts as an Oracle. The scheduler queries this manager via specific interfaces (e.g., `trustManager.isTrusted(vmId)`) rather than embedding logic in the scheduler itself.

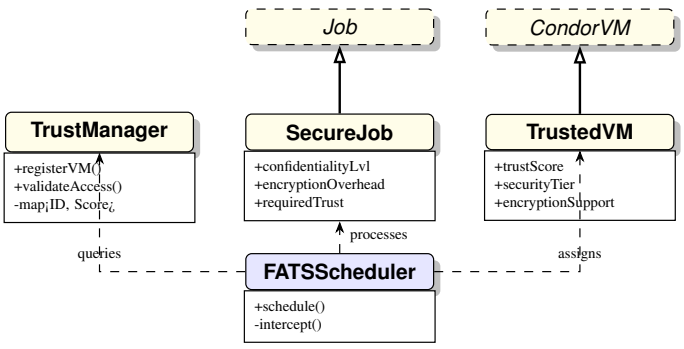


Fig. 1. UML class diagram of the trust-aware WorkflowSim extension. It illustrates the inheritance hierarchy where `SecureJob` extends the base `Job` class, and `TrustedVM` extends `CondorVM`, enabling security attributes to be handled natively by the simulation engine.

D. Execution Flow Modifications

In standard WorkflowSim, the `WorkflowScheduler` maps jobs based on performance. In our extended framework, we inject a "Security Interceptor" that allows *any* custom logic to be applied:

- **Task Parsing:** Workflows are parsed and generic sensitivity labels are assigned.
- **Interceptor Hook:** The `schedule()` method calls a `validateSecurity()` hook.
- **Overhead Injection:** A simulation penalty is added to tasks requiring encryption.
- **Feedback Loop:** Upon task completion, the `TrustManager` receives a signal to update internal state.

E. Algorithmic Framework and Complexity Analysis

The core scheduling logic, orchestrated by the `FATSScheduler`, integrates the fragmentation and trust-aware filtering phases. The time complexity of this integration is critical for ensuring low-latency decision-making in dynamic fog environments.

1) Algorithm Workflow: The scheduling process follows three main phases:

- **Initialization Phase:** The DAG is traversed to compute task levels and priorities. The initial trust scores for all nodes are loaded from the `TrustManager`.
- **Filtering Phase:** For each ready task t_i, the candidate set of resources V is filtered to create a secure subset $V_{secure} \subseteq V$ such that $\forall v_j \in V_{secure}, T_j \geq T_{req}(S_i)$.
- **Scoring Phase:** Remaining candidates are scored based on execution time and cost, and the task is assigned to the optimal node.

2) Complexity analysis: Let N be the number of tasks in the workflow and M be the number of available fog resources.

a) Task provisioning overhead: The standard WorkflowSim provisioning algorithm operates in $O(N \cdot M)$. Our extension introduces the filtering step. Checking the trust constraint for M resources takes $O(M)$. Since this is repeated for every scheduling decision, the base complexity remains $O(N \cdot M)$.

However, sorting the filtered list V_{secure} (where $|V_{secure}| \leq M$) takes $O(M \log M)$. Thus, the total complexity per scheduling round is:

$$T_{sched} = O(N \cdot (M + M \log M)) \approx O(N \cdot M \log M) \quad (11)$$

This logarithmic factor is acceptable given that M (number of fog nodes) is typically much smaller than N (number of tasks) in typical scientific workflows ($N \approx 1000, M \approx 20 - 50$). The overhead is negligible compared to the network latency involved in task offloading.

V. Implementation Details

The implementation is built on Java (JDK 1.8) integrating with WorkflowSim 1.1.0. This section highlights the critical code segments responsible for the security logic.

A. The TrustManager

The `TrustManager` is the heart of the security extension. It allows for dynamic updates to node reputation. We implemented a generic Beta Reputation System interface capable of modeling successful (α) and failed (β) interactions.

The update logic follows the standard expectation model:

$$T_{new} = \frac{\alpha + 1}{\alpha + \beta + 2} \times \delta_{history} \tag{12}$$

where, T_{new} is the updated trust score and $\delta_{history}$ is a decay factor for aging past interactions.

Listing 1: TrustManager and Secure Scheduling Logic

```
public class TrustManager {
    private Map<Integer, Double> vmTrustScores;
    // ... constructor omitted ...

    public void registerVM(int vmId, String tier) {
        double initialTrust = 0.80;
        if (tier.equals("MIST")) initialTrust = 0.95;
        if (tier.equals("CLOUD")) initialTrust = 0.75;
        vmTrustScores.put(vmId, initialTrust);
    }
    public boolean validateAccess(int vmId, int
         sensitivity) {
        double score = vmTrustScores.getOrDefault(vmId,
             0.5);
        if (sensitivity == 3 && score < 0.90) return false;
        // ... additional checks ...
        return true;
    }
}

// --- FATSScheduler: Secure Job Scheduling ---
protected void scheduleSecureJob(SecureJob job) {
    if (job.isSensitive()) {
        List<SecureJob> fragments =
            fragmentationEngine.split(job);
        for (SecureJob fragment : fragments) {
            TrustedVM targetVM = trustManager
                .findVM(fragment.getRequiredTrust());
            submitToEngine(fragment, targetVM);
        }
    } else {
        super.scheduleJob(job);
    }
}
```

B. Secure Job Scheduling

The scheduler logic was extended to support "Split Execution" capabilities. This allows testing of algorithms that break tasks into sub-components (e.g., Secure vs. Public fragments) without hardcoding the split ratio. The scheduling interceptor method `scheduleSecureJob()` shown in Listing 1 demonstrates how the `FATSScheduler` delegates trust validation to the `TrustManager` and applies fragmentation when a task is sensitive.

C. Overhead Modeling

To ensure the simulation reflects the "cost" of security, we implemented a data-size-dependent delay. This prevents the simulator from falsely reporting that secure scheduling is "free".

$$T_{overhead} = \frac{D_{size}}{B_{enc}} \times \alpha$$

where, D_{size} is the data size, B_{enc} is the encryption throughput of the VM, and α is a processing coefficient.

D. Configuration and Usage

To maximize adoption, distinct configuration files are used to separate security parameters from infrastructure setup.

1) Security configuration: The framework introduces `security.properties`, allowing researchers to tune crypto-overhead without recompiling:

```
security.encryption.overhead=5.2 # ms per KB
security.trust.decay_factor=0.01
security.trust.reward_factor=0.005
```

VI. Case Study: Secure Fog Workflow Scheduling

To demonstrate the functionality and extensibility of the proposed framework, we conducted a comprehensive case study. This section details the experimental setup, workflow configurations, and performance metrics, followed by scenario-based analysis.

A. Experimental Setup

The simulation environment was configured to model a realistic fog computing hierarchy consisting of three distinct tiers: Mist, Fog, and Cloud. Table III outlines the core simulation parameters used in this study. The experiments were conducted on a standard Windows 10 workstation with an Intel i7-8550U processor (1.80 GHz) and 16 GB of RAM.

TABLE III. Simulation Parameters

Parameter	Value description
Platform	WorkflowSim 1.1.0 on JDK 1.8
Datacenters	3 (Mist, Fog, Cloud)
VM Types	Mist (500 MIPS), Fog (1000 MIPS), Cloud (2500 MIPS)
Default Trust Levels	Mist (0.95), Fog (0.85), Cloud (0.75)
Task Confidentiality	High (3), Medium (2), Low (1)
Encryption Overhead	5.2 ms/MB
Scheduling Policy	Dynamic Trust-Aware Minimum Completion Time

B. Workflow Characteristics

Scientific workflows are characterized by complex data dependencies and varying computational requirements. We sourced five standard scientific workflows from the Pegasus Workflow Generator [32], [33]: Montage, CyberShake, Epigenomics, Inspiral, and Sipht.

Table IV summarizes the characteristics of these benchmark workflows. Each workflow type represents a different application domain (e.g., Montage for astronomy, CyberShake for earthquake science).

TABLE IV. SCIENTIFIC WORKFLOW BENCHMARKS

Name	Tasks	Jobs	Avg MI/Task	Data (MB)
Montage	100	201	12,500	3.2
CyberShake	100	201	45,000	102.4
Epigenomics	100	201	28,000	55.8
Inspiral	100	201	62,000	12.6
Sipht	100	201	18,500	8.4

C. Performance Metrics

The simulator was extended to calculate specific metrics relevant to secure fog computing. These metrics enable researchers to monitor the trade-offs between performance and security.

- Makespan (seconds): The total execution time of the workflow, from the submission of the first task to the completion of the last. It includes processing time, data transfer time, and security overhead.
- Total Cost (USD): The cumulative cost incurred by executing tasks on leased Cloud/Fog VMs. Cost models differ by tier (e.g., Cloud is cheaper per instruction but incurs higher latency).
- Security Score: A normalized value $(0.0 - 1.0)$ representing the average trust level of all VMs utilized during execution.

$$S_{score} = \frac{1}{n} \sum_{i=1}^{n} T(VM_i) \times C(task_i) \tag{13}$$

where, $T(VM)$ is the node's trust score and $C(task)$ is the confidentiality weight.

- Failure Rate: The percentage of tasks that failed due to trust violations or node unavailability.

D. Scenario 1: Functional Verification

We simulated a "Malicious Insider" scenario involving 100 tasks to verify the enforcement capability. We manually configured 2 Fog nodes (IDs 4 and 5) to have a Trust Score of 0.4, simulating compromised devices.

- Setup: 100 Tasks (50 High Sensitivity, 50 Low Sensitivity).
- Expected Behavior: The simulator should block all 50 High Sensitivity tasks from the compromised nodes.
- Observed Result: Analysis of the simulation logs confirmed that 0 High Sensitivity tasks were scheduled on the low-trust nodes. They were either queued or rerouted to Mist nodes.

Fig. 2 shows a snippet of the console output generated during this validation test. The logs clearly indicate the `TrustManager` rejecting the assignment and the Scheduler triggering a reroute event.

```
[INFO] 10.0: TrustManager: VM #4 (Fog Tier) Trust Score updated to 0.45 due to
       previous failure.
[INFO] 10.1: WorkflowScheduler: Attempting to schedule Task #15 (High
       Confidentiality) to VM #4.
[WARN] 10.1: TrustManager: ACCESS DENIED. VM #4 Trust (0.45) < Required (0.90).
[INFO] 10.1: WorkflowScheduler: Security Violation detected. Rerouting Task #15...
[INFO] 10.2: WorkflowScheduler: Task #15 successfully assigned to VM #1 (Mist Tier,
       Trust 0.96).
[INFO] 10.5: CloudSim: Job #15 started execution on VM #1.
```

Fig. 2. Console trace demonstrating dynamic security enforcement.

E. Scenario 2: Performance Impact Analysis

To demonstrate the simulator's ability to model trade-offs, we ran the Montage workflow under varying security constraints. The results shown in Table V illustrate how enforcing strict security policies impacts overall makespan and cost.

TABLE V. IMPACT OF SECURITY POLICIES ON WORKFLOW METRICS

Policy	Makespan (s)	Cost ($)	Security Score
None (Baseline)	428.5	1.25	0.65 (Low)
Trust-Aware	512.3	1.45	0.88 (High)
Encrypted	545.1	1.62	0.95 (V. High)

As expected, enabling trust-awareness increases the makespan by approximately 19% due to rerouting tasks away from powerful but untrusted nodes. The encryption overhead further adds a 6% delay. This data confirms that the simulator accurately reflects the performance penalties associated with secure computing, providing a valuable testbed for optimization research.

F. Simulation Performance Overhead

Adding security checks introduces computation overhead to the simulation engine itself. We compared the "Wall Clock Time" (how long the simulation took to run on the host PC) for Standard WorkflowSim vs. our Trust-Aware Extension (see Table VI).

TABLE VI. SIMULATION ENGINE OVERHEAD ANALYSIS

Workload Size	Standard Time	Extended Time	Overhead %
100 Tasks	450 ms	482 ms	+7.1%
500 Tasks	2100 ms	2250 ms	+7.1%
1000 Tasks	4500 ms	4810 ms	+6.8%

The overhead introduced by the trust validation and encryption modeling remains consistently below 7.1%, which is negligible compared to the simulated task execution times. This confirms that the security extension does not impose a prohibitive computational burden on the simulation engine itself.

G. Limitations

While the proposed framework demonstrates functional correctness and acceptable overhead, several limitations should be acknowledged. First, the network latency model is simplified and does not capture real-world phenomena such as jitter, packet loss, or dynamic routing. Second, the current trust degradation model is triggered only by task execution failures; it does not account for reputation attacks, collusion among malicious nodes, or time-based trust decay in the absence of

interactions. Third, the framework has been validated with up to 1000 tasks and approximately 50 fog nodes; scalability under hundreds of concurrent fog nodes with real-time trust updates requires further investigation. Fourth, validation is based on trace analysis and functional verification rather than comparison with physical fog testbeds. This approach is consistent with the validation methodology of established simulation tools such as CloudSim [14] and iFogSim [17], which similarly rely on scenario-based analysis to demonstrate correctness. Integration with real-world testbed data is identified as an important direction for future work.

VII. Discussion

The experimental results demonstrate that the proposed trust-aware extension to WorkflowSim successfully enforces security constraints while introducing quantifiable performance trade-offs. The 19% increase in makespan observed under trust-aware scheduling is attributable to the rerouting of sensitive tasks from high-capacity but low-trust fog nodes to lower-capacity but trusted mist nodes. This trade-off is consistent with findings reported in the literature on security-aware scheduling [27], where security enforcement typically incurs a 15–25% overhead on task completion time.

Unlike existing approaches that hard-code trust checks within scheduling algorithms, our modular `TrustManager` architecture enables researchers to independently evaluate and compare different trust models (e.g., Beta Reputation, EigenTrust, blockchain-based) without modifying the core scheduling logic. This separation of concerns represents a significant improvement over ad-hoc implementations [34] and promotes reproducibility in fog computing security research.

The simulation engine overhead of less than 7.1% confirms that the security extension is lightweight and suitable for large-scale simulation studies. This low overhead is achieved through the use of $O(1)$ trust lookups via hash-map-based storage in the `TrustManager`, avoiding the iterative computation required by global trust algorithms.

VIII. Conclusion

This study presented the design and implementation of a trust-aware extension to WorkflowSim, addressing the critical gap in existing fog computing simulation tools that lack native support for security modeling. The key contributions include: 1) an extended class hierarchy (`SecureJob`, `TrustedVM`) that integrates trust and confidentiality as first-class simulation attributes; 2) a modular `TrustManager` component that decouples security logic from scheduling logic, enabling interchangeable trust models; and 3) an encryption overhead model that accurately reflects the computational cost of security in simulation results.

Experimental validation confirmed that the framework correctly enforces security policies, blocking all high-sensitivity tasks from compromised nodes. The performance analysis revealed a 19% makespan increase and 16% cost increase under trust-aware scheduling compared to the baseline, quantifying the security-performance trade-off that is often overlooked in existing simulators. The simulation engine overhead remains below 7.1%, confirming the practical viability of the extension for large-scale studies.

The scientific value of this work lies in providing the fog computing research community with a standardized, extensible simulation platform for evaluating security-aware scheduling algorithms. By open-sourcing the framework and adopting modular design principles, we enable reproducible research and facilitate fair comparison across different trust models and scheduling strategies. Future work will focus on integrating real-world network trace data, supporting blockchain-based trust models [35], and validating the framework against physical fog testbed deployments.

References

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[2] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[3] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[4] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[5] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[6] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," in *Proc. 1st Edition of the MCC Workshop on Mobile Cloud Computing*, pp. 13–16, 2012.

[7] P. Habibi, M. Farhoudi, S. Kazemian, S. Khorsandi, and A. Leon-Garcia, "Fog computing: A comprehensive architectural survey," *IEEE Access*, vol. 8, pp. 69105–69133, 2020.

[8] M. Mukherjee, R. Matam, L. Shu, L. Maglaras, M. A. Ferrag, N. Choudhury, and V. Kumar, "Security and privacy in fog computing: Challenges," *IEEE Access*, vol. 5, pp. 19293–19304, 2017.

[9] S. Yi, C. Li, and Q. Li, "A survey of fog computing: Concepts, applications and issues," in *Proc. Workshop on Mobile Big Data (Mobidata)*, pp. 37–42, 2015.

[10] A. V. Dastjerdi and R. Buyya, "Fog computing: Helping the Internet of Things realize its potential," *Computer*, vol. 49, no. 8, pp. 112–116, 2016.

[11] I. Stojmenovic and S. Wen, "The fog computing paradigm: Scenarios and security issues," in *Proc. Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 1–8, 2014.

[12] J. Ni, K. Zhang, X. Lin, and X. S. Shen, "Securing fog computing for Internet of Things applications: Challenges and solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 601–628, 2018.

[13] R. Roman, J. Lopez, and M. Mambo, "Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges," *Future Generation Computer Systems*, vol. 78, pp. 680–698, 2018.

[14] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and 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.

[15] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616, 2009.

[16] W. Chen and E. Deelman, "WorkflowSim: A toolkit for simulating scientific workflows in distributed environments," in *Proc. IEEE 8th Int. Conf. E-Science*, pp. 1–8, 2012.

[17] H. Gupta, A. V. Dastjerdi, S. K. Ghosh, and R. Buyya, "iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments," *Software: Practice and Experience*, vol. 47, no. 9, pp. 1275–1296, 2017.

[18] R. Mahmud, R. Kotagiri, and R. Buyya, "Fog computing: A taxonomy, survey and future directions," in *Internet of Everything*, Springer, Singapore, pp. 103–130, 2018.

[19] C. Puliafito, E. Mingozzi, F. Longo, A. Puliafito, and O. Rana, "Fog computing for the Internet of Things: A survey," *ACM Transactions on Internet Technology*, vol. 19, no. 2, pp. 1–41, 2019.

[20] C. Sonmez, A. Ozgovde, and C. Ersoy, "EdgeCloudSim: An environment for performance evaluation of edge computing systems," *Transactions on Emerging Telecommunications Technologies*, vol. 29, no. 11, e3493, 2018.

[21] C. Mechalikh, H. Taktak, and F. Moussa, "PureEdgeSim: A simulation framework for performance evaluation of cloud, edge and mist computing environments," *Computer Science and Information Systems*, vol. 18, no. 1, pp. 43–66, 2021.

[22] T. Qayyum, A. W. Malik, M. A. Khan Khattak, O. Khalid, and S. U. Khan, "FogNetSim++: A toolkit for modeling and simulation of distributed fog environment," *IEEE Access*, vol. 6, pp. 63570–63583, 2018.

[23] A. Josang and R. Ismail, "The beta reputation system," in *Proc. 15th Bled Electronic Commerce Conference*, 2002.

[24] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The EigenTrust algorithm for reputation management in P2P networks," in *Proc. 12th Int. Conf. World Wide Web*, pp. 640–651, 2003.

[25] M. Nitti, R. Girau, and L. Atzori, "Trustworthiness management in the Social Internet of Things," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1253–1266, 2014.

[26] Z. Yan, P. Zhang, and A. V. Vasilakos, "A survey on trust management for Internet of Things," *Journal of Network and Computer Applications*, vol. 42, pp. 120–134, 2014.

[27] T. Wang, G. Zhang, A. Liu, M. Z. A. Bhuiyan, and Q. Jin, "A secure IoT service architecture with an efficient balance dynamics based on cloud and edge computing," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4831–4843, 2019.

[28] M. A. Khan and K. Salah, "IoT security: Review, blockchain solutions, and open challenges," *Future Generation Computer Systems*, vol. 82, pp. 395–411, 2018.

[29] H. Topcuoglu, S. Hariri, and M.-Y. Wu, "Performance-effective and low-complexity task scheduling for heterogeneous computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, no. 3, pp. 260–274, 2002.

[30] Z. Zhu, G. Zhang, M. Li, and X. Liu, "Evolutionary multi-objective workflow scheduling in cloud," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 5, pp. 1344–1357, 2016.

[31] L. F. Bittencourt, J. Diaz-Montes, R. Buyya, O. F. Rana, and M. Parashar, "Mobility-aware application scheduling in fog computing," *IEEE Cloud Computing*, vol. 4, no. 2, pp. 26–35, 2017.

[32] S. Bharathi, A. Chervenak, E. Deelman, G. Mehta, M.-H. Su, and K. Vahi, "Characterization of scientific workflows," in *Proc. 3rd Workshop on Workflows in Support of Large-Scale Science*, pp. 1–10, 2008.

[33] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, and K. Wenger, "Pegasus, a workflow management system for science automation," *Future Generation Computer Systems*, vol. 46, pp. 17–35, 2015.

[34] Z. Xiao and Y. Xiao, "Security and privacy in cloud computing," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 2, pp. 843–859, 2013.

[35] S. S. Gill, S. Tuli, M. Xu, I. Singh, K. V. Singh, D. Lindsay, S. Tuli, D. Smiber, M. Usman, R. Buyya, et al., "Transformative effects of IoT, Blockchain and Artificial Intelligence on cloud computing: Evolution, vision, trends and open challenges," *Internet of Things*, vol. 8, 100118, 2019.

User-Centered Ergonomic Design of a Portable EEG System for Evaluating Dolphin-Assisted Therapy in Children with Neurodevelopmental Disorders

Brenda Lorena Flores Hidalgo[1], Urim Rarael Pérez Bernal[2],
Jesús Jaime Moreno Escobar[3], Hugo Quintana Espinosa[4], Ana Lilia Coria Páez[5]
Escuela Superior de Ingeniería Mecánica y Eléctrica, Zacatenco, Instituto Politécnico Nacional, México[1,2,4]
Departamento de Sistemas, División de Ciencias Básicas e Ingeniería,
Universidad Autónoma Metropolitana, Azcapotzalco, México[1]
Centro de Investigación en Computación, Instituto Politécnico Nacional, México[3]
Escuela Superior de Comercio y Admistración, Tepepan, Instituto Politécnico Nacional, México[5]

***Abstract*—Dolphin-Assisted Therapy (DAT) is a therapeutic alternative used in comprehensive neurorehabilitation, aimed at improving the quality of life and social integration of patients with neurodevelopmental disorders through interaction with dolphins. However, the lack of defined standards makes it difficult to objectively evaluate its effectiveness. In this context, it was identified that the electroencephalographic device used to record EEG signals before, during, and after the sessions (based on a TGAM1 sensor) presented both functional and ergonomic limitations, as its readjustment interrupted the therapy sequence and caused discomfort in patients. To address this issue, the present work proposes the redesign of the device, incorporating ergonomic criteria to improve comfort, stability, and ease of use, without compromising the quality of the signal acquired by the TGAM1 sensor. The main objective was to optimize the user experience while also enabling the collection of more consistent and reliable data for analysis. The results show a significant improvement, achieving a 51% increase in signal retention. This made it possible to recover approximately 36 additional seconds of high-quality neural data per hour of therapy, thanks to more continuous and accurate EEG acquisition during interactions with dolphins. Furthermore, patient evaluations indicated greater acceptance of the device, highlighting improvements in comfort and weight perception compared to the previous version, thereby validating both its functionality and ergonomic design.**

***Keywords*—User-centered design; ergonomics; portable EEG system; Dolphin-Assisted Therapy; neurodevelopmental disorders; 3D printing; TGAM1**

I. Introduction

Dolphin-Assisted Therapy (DAT) is an alternative therapeutic method in which the well-feeling of children with various neurodivergent conditions is pursued via interaction between humans and dolphins, typically pregnant or lactating females from the bottlenose species. These therapies act as adjuncts to standard treatments that aim to promote healing and wellness [1]. Some studies have demonstrated that dolphin-assisted treatments accelerated motor skills, improved motivation for learning, benefitted on cardiovascular and respiratory efficiency, increased the flexibility of movement, help to reduce pain and helped in balance [2]. To date, DAT has been studied mainly to prove its efficacy [3].

Currently, DAT lacks established standards. Thus, a standardized approach is necessary to ensure that the therapeutic advantages of FIT are fully realized. According to Rutherford [4], standardization enables the ends of treatment efficacy evaluation. This enables better comparisons of results from different treatments, allowing professionals to determine the most effective techniques for specific diseases. In the case of DAT, for example, one measurement of its efficacy can be made by looking at variation in neurodivergent children's brain waves (as measured through electroencephalographic [EEG] signals) before, during, and immediately after the treatment with DAT. Depending on cerebral cortex region, the 10/20 system [5] defines a set of specific points on the scalp where several electrodes must touch to measure an EEG. This enables the measurement of electrical impulses generated by neurons to study human cognitive activity. Standardization of DAT is a problem, because the approach often differs widely in terms of method, frequency and duration of sessions, training used and species/type of dolphins [6], [7], [8], which can produce great variation in results. Therefore, the existence of measurement methods and instruments to capture EEG signals before, during, and after each session is essential to analyze the effectiveness of DAT.

Repositioning the electroencephalographic apparatus recording EEG signals before, after and during DAT represents a considerable challenge faced in the standardization of DAT.

In this way, the main goal is not to improve on the inherent performance of the sensor which will always be fixed by design, but rather, to obtain a match with its theoretical performance by simulating to some extent the favorable (or ideal) operating conditions where the sensor was initially proven.

In addition, it was determined that the moments of placement and withdrawal of devices are different between the types of conditions of the child being evaluated with DAT, so much so that they interfere with the performance order in which operations were carried out throughout the sessions, generating some discomfort for its use [9].

Research in EEG focuses both on understanding human cognition and on the use of brain signals to interact with the external environment through brain–computer interfaces (BCI).

A large portion of this research is conducted in medical or academic institutions [10].

There are multiple potential applications outside the laboratory environment, including prosthetic control, communication for individuals without motor function, therapeutic rehabilitation, and video games [11]. However, implementing EEG in real-world environments presents several challenges, particularly in uncontrolled contexts such as DAT. In these scenarios, expert personnel are not always available, electrode caps are often complex, and their placement requires a considerable amount of time. Furthermore, considering that patients are often neurodivergent children, it is essential that the device be comfortable and not cause discomfort during use.

In this way, the main objective on signal retention in an aquatic therapeutic environment, where conventional EEG devices face ergonomic challenges from patient movement and signal interference from the physical setting, addresses a specific and underexplored design problem, and the quantified improvement in signal retention provides a concrete metric of engineering value.

Therefore, the primary goal is the maintenance of the signal in a therapeutic water setting; traditional EEG devices are ergonomically complicated with behaviors such as patient movement and physical environment signals. It addresses a concrete design problem that has thus far been underdiscussed and the measured improvement in signal retention quantifies an engineering value.

Ergonomics is the applied science that studies interactions between human beings and other elements of a system. This is done by applying theories, principles, data and design methods whose end aim is improving human well being and system performance [12], [13], but only few studies include ergonomic aspects in the design of electroencephalographic devices [14], [15]. The current research describes the re-design of an electroencephalographic device utilized for EEG signals acquisition incorporating specific ergonomic concepts. In this redesign, we aim to create an instrument that is not only comfortable, functional, and aesthetically appealing for the patient, but also decreases applicable downtime identified in the DAT standardization [9], so as to maximize stable data collection which will allow multi-site analysis protocols or cross-instrument comparison of study conclusions as well as support the development of new evidence-based therapeutic strategies.

The structure of this study is as follows: Section II provides the fundamentals of electroencephalography and ergonomics that made it possible to redesign the electroencephalographic equipment employed pre-, peri- and post-DAT. Section III illustrates the methodology applied in this study. Section IV describes the Redesign and Development Process of Electroencephalographic Device. Section V provides the results and discussion of analysis. Lastly, the conclusion and future work is given in Section VI.

II. Theoretical Framework

A. Electroencephalography

Electroencephalogram (EEG) signals can be measured non-invasively through electroencephalography, which is widely used to obtain electrical impulses generated from human cognitive actions [16]. For research purposes this technique has been utilized for neuroscience, psychology etc.; designing brain-computer interface applications for diagnosis of several disorders including epilepsy, Attention-Deficit/Hyperactivity, etc. [17], [3]. EEG enables the recording of electrical potentials via several electrodes touching the scalp at designated locations on the cerebral cortex, as defined by an international system known as 10/20 system (see Fig. 1; [5]).

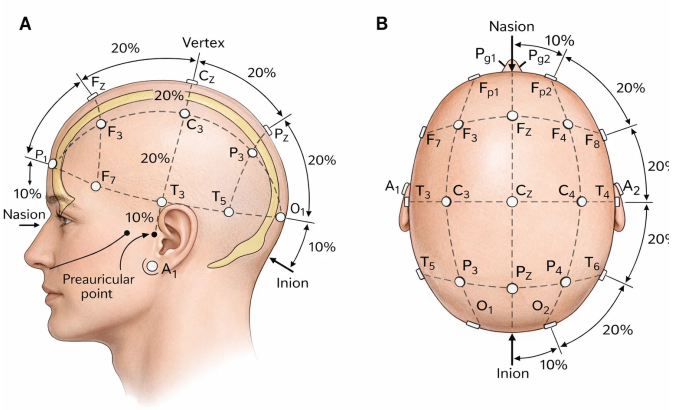


Fig. 1. International 10-20 system for the hydration of extracranial electrodes (Fp: Fronto-polar; F: Frontal; C: Central; P: Parietal; T: Temporal and O: Occipital), based on [5].

Electroencephalography has several advantages over other imaging and behavioral observation methods such as Magnetoencephalography and Magnetic Resonance Imaging, including its excellent temporal resolution and low maintenance cost [3]. Dry electrodes do not require the application of an electrolyte, making their manipulation and cleaning easier. In addition, they enhance ergonomics and comfort designed especially for non-expert users [18]. At the moment, new materials are enabling the production of wireless, miniaturized systems for enhanced applicability [15]. If established this technology adaptation and enhancement might allow its employment beyond scientific labs or hospital, and possible studies around real world situations, domestic health usage as well as the analysis, treatment and supervision of patients with neurodevelopmental problems [19], [20]. However, the implementation of EEG systems in real-world environments presents several challenges, particularly in uncontrolled contexts such as DAT. In these scenarios, specialized personnel are not always available, electrode caps are often complex to handle, and their placement requires a considerable amount of time. Furthermore, since the patients are often neurodivergent children, it is essential that the device provide comfort and not cause discomfort during its use.

Therefore, the development of systems that are easier to use and allow more precise electrode placement could significantly expand real-world BCI applications [21], [22]. In this regard, several studies recommend the design of more ergonomic EEG devices that better adapt to the diversity of head shapes and sizes, as well as to anatomical reference points [23]. To improve usability and accuracy, devices should be designed so that the electrodes are positioned as close as possible to

their ideal locations according to the 10–20 system, while also ensuring repeatability between measurement sessions. This aspect is particularly relevant in DAT sessions, since, as previously mentioned, this field still lacks fully defined standards for data measurement and analysis. Therefore, having a device capable of adapting to the specific conditions of the environment in which this therapy takes place is essential for advancing research in this area.

B. Ergonomics

Ergonomics is a scientific fields devoted to the study of interactions among humans and their elements of a system. It is done by applying theories, principles, data and design methods to enhance human well-being and the overall system performance [12], [13]. Product ergonomics studies the needs of users and the conditions surrounding use. Taking into account ergonomic factors enables product design which allows adapting a product to the features of end users to ensure direct efficiency, safety and productivity in its operation so that it does not cause health problems for people [24].

Disparities between individuals such as gender, race, and abilities can cause product design to fall short. When a device fulfills both the functional and emotional needs of its users [25], it becomes more attractive. As previously mentioned, much of the research on brain–computer interfaces (BCIs) has focused on different technical aspects, such as data acquisition, signal processing, the theoretical foundations of BCIs, as well as the development of conceptual applications or innovative prototypes [26], [27]. Although these research lines are essential to ensure the reliability of BCI technology, studies focused on the ergonomics of these devices remain limited. In particular, further research is needed to analyze how factors such as comfort, usage limits, and the physical characteristics of users influence the suitability of BCI systems for implementation in experimental environments and real-world applications. Such studies are essential to fully exploit the potential of these technologies in a wide variety of contexts.

Despite the progress achieved, the benefits of BCI systems are often limited by several factors, including low signal reliability [26], [28], fatigue associated with prolonged device use, and, in some cases, difficulties related to usability. These limitations raise questions about the suitability of these devices for both research purposes and practical applications. In the case of physiological measurement devices such as EEG systems, ergonomic design becomes particularly important, since the equipment must remain in direct contact with the user's head during signal acquisition. For this reason, it is essential to consider factors such as weight distribution, the pressure exerted by the electrodes, ease of placement, and the ability to adapt to different cranial morphologies [10]. An appropriate design not only improves user comfort but also contributes to obtaining more stable and reliable recordings.

In the context of DAT, the design of the EEG device must also consider the specific conditions of the environment in which the sessions take place. Factors such as user mobility, interaction with an environment close to water, and the need to maintain electrode stability during the therapeutic session require the development of specialized ergonomic solutions. Therefore, designing a device that combines comfort, stability, and ease of use is essential to ensure both the quality of electroencephalographic recordings and an appropriate user experience during therapeutic sessions.

C. Electronics Behind EEG

1) Acquisition: The TGAM1 (ThinkGear ASIC Module) is a single-channel low-power biosensor chip designed by NeuroSky for embedded electroencephalography. The module combines the TGAT chip and is intended to be used in portable and portable BCI devices, as shown in Fig. 2. The TGAM1 works in a voltage range of 2.97 V to 3.63 V and consumes a maximum current of 15 mA @ 3.3 V, leading to very low power consumption as necessary for battery operated applications like the EEG for Dolphin-Assisted Therapy application. It shows an extra-small size of $27.9mm \times 15.2mm \times 2.5mm$ and weighs only around 130 mg, making it suitable for use in headset applications where the module will not interfere with user comfort. It has a 12 bit analog to digital converter which supports 512 Hz sampling rate, and it can capture EEG signals in the range of 0.5 Hz to 100 Hz. The module needs three dry contact electrodes, an electrode for EEG signal registration, and a second as a reference; the solution is to connect the third electrode as the ground. The TGAM1 also has 4 kV of contact discharge and 8 kV of air discharge resistance for built-in electrostatic discharge protection, allowing it to operate reliably under a variety of environmental conditions.

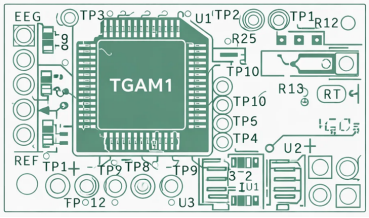


Fig. 2. ThinkGear ASIC Module: Single-channel low-power biosensor chip designed by NeuroSky for embedded electroencephalography.

TGAM1 is a UART serial device with configurable baud rates of 1200, 9600, and 57600 bps; the factory setting for normal data exchange is after booting, having it fixed at the value of 9600 bps. The module also has a built in notch filter to remove power line interference, which can be switched by connecting *the M* pin: connecting it to Vcc enables 60 Hz filtering for countries with 60 Hz AC, such as Mexico or USA; while connecting it to ground filters at 50 Hz for those using 50 Hz AC like Spain or France. The module also has configuration pads (B_0 and B_1) on the back that settle the baud rate and data output mode for a permanent configuration. The TGAM1 is designed to output raw EEG data at 512 Hz and processed parameters, including proprietary NeuroSky Attention and Meditation meters scaled from 0 to 100 percent, along with efficient eye blink detection. These processed outputs are particularly useful for therapeutic applications, as they offer real-time indicators of patient cognitive states without a need for sophisticated signal processing algorithms on the host device.

2) Wireless communication: The first device you will need for a wireless EEG system used in Dolphin-Assisted Therapy is the HC-05 Bluetooth module, Fig. 3, which allows the patient

to transmit their brain activity data over the air to a monitoring station without being tethered by heavy wires. In addition, The TGAM1 EEG sensor captures the neural signals needed to transmit with this module through the HC-05 Interface wirelessly in real-time while children have free interaction with dolphins.

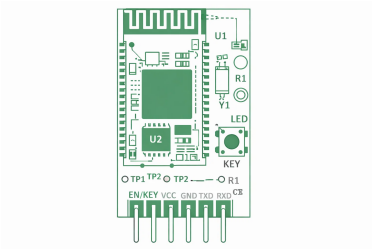


Fig. 3. HC-05 Bluetooth serial module: indicating it's onboard PCB antenna, its main controller ICs and some supporting components.

In this way, HC-05 is a traditional Bluetooth module based on the Bluetooth specification V2.0 and Enhanced Data Rate (EDR) for transparent wireless serial communication. It works on the 2.4 GHz ISM frequency band with GFSK modulation, and CSR's BC417 chip is used to provide reliable wireless connectivity. It has an RF transmit power of +4dBm and reception sensitivity of -85 dBm, allowing communication up to 10 meters in open space. The HC-05 measures just $27mm \times 13mm \times 2mm$ and weighs about 3.7 grams, and includes a PCB antenna with its interface covered by a connector design featuring the following pin layout on a 2.54 mm pitch header: STATE, RXD, TXD, GND, VCC, and EN. This module runs with a supply voltage of 3.6 V to 6 V and consumes 40 mA when pairing and 8 mA when connected. An onboard LED indicates the module's operational status: fast blinking for no connection, slow blinking for AT command mode, and double blinking for successful connection.

The HC-05 communicates via a UART interface, with adjustable baud rates of 4800, 9600, 19200, 38400, 57600, 115200, 230400, 460800, 921600, and 1382400 bps, with factory default settings of 38400 bps, 8 data bits, 1 stop bit, and no parity. The module has two operational working modes: 1) at command response mode for AT command configuration and automatic connection mode for data transmission, it can work as Master, Slave or Loopback roles; and 2) AT_COMMAND_MODE: This mode can be accessed by powering on the module while holding down the enable button, which will make the LED blink slowly during start up (The default baud rate for AT command mode is 38400 bps) Some key configurable parameters (AT+NAME= for Name, AT+ROLE=0 for Slave and 1 for Master, AT+PSWD; default is "1234" or "0000", etc.). It can all be done via serial port (AT+UART). The STATE pin shows whether a connection is active or not; it will output HIGH when connected and LOW when disconnected, allowing the microcontroller to track the status of the connection. Supported features, the module supports power-on auto-connect last paired device and adaptive frequency hopping to minimize interference.

III. Methodology

The methodology (Fig. 4) used for the redesign of the electroencephalographic device that retrieves the EEG signals before, during and after DAT with ergonomic concepts in order to bring comfort, functionality and beauty for the patient. Applied methodology: Soft Systems Methodology from the study proposed by Coria et al. [9] to demonstrate the B2 Model of relevant activities for the development of DAT-specific measurement tools.

Stage 1: In this stage, the elements that make up the electroencephalographic device were developed using Computer-Aided Design (CAD) tools. The objective was to generate a detailed digital representation of the device at a 1:1 scale, allowing precise visualization of the shape, dimensions, and arrangement of each component. During this phase, structural aspects such as the support base, electrode placement points, adjustment mechanisms, and fastening elements required to keep the device stable during use were defined. Additionally, the digital modeling made it possible to anticipate potential interferences between components and carry out preliminary adjustments before the physical prototype was manufactured.

Stage 2: In this stage, the main ergonomic factors that needed to be considered for the redesign of the device were identified and analyzed. These factors included aspects such as system weight distribution, the pressure exerted by the electrodes on the scalp, ease of placing and removing the device, and adaptability to different cranial morphologies. Criteria related to user comfort, system stability during movement, and interaction with the therapeutic environment were also evaluated. This analysis made it possible to establish design guidelines aimed at improving the user experience and ensuring that the device could be used during therapy sessions without causing discomfort or interfering with EEG signal acquisition.

Stage 3: In this stage, the results obtained from Stages 1 and 2 were integrated through analysis sessions and brainstorming activities, in which different design alternatives were evaluated to meet both the technical requirements and the previously identified ergonomic criteria. Based on this integration, the device design was developed using SolidWorks, which enabled detailed three-dimensional modeling of the system. During this phase, the final dimensions of the device, the precise location of the electrodes, the adjustment mechanisms, and the structural elements required to ensure stability, comfort, and functionality during use were defined.

Stage 4: Once the digital design of the device was completed, the physical prototype was manufactured using 3D printing. This process allowed a tangible version of the device to be produced quickly and at low cost, facilitating early evaluation of the proposed design. The use of additive manufacturing technologies made it possible to accurately reproduce the geometries created in the CAD model, while also allowing rapid modifications if necessary. The prototype was printed in resin to obtain a lightweight and resistant structure suitable for initial functionality testing.

Stage 5: In this step, various tests were carried out to evaluate the performance of the developed prototype. These tests included verifying the overall functionality, user comfort during use, system stability when placed on the user's head, and the proper distribution of the components' weight.

Additionally, the ease of placing and removing the device was analyzed, as well as its ability to maintain the position of

the electrodes at the anatomical points required for capturing EEG signals. The results obtained at this stage made it possible to identify strengths of the proposed design and potential areas for improvement.

Stage 6: Finally, in this stage the results obtained during the tests performed in Stage 5 were analyzed in order to identify possible modifications or adjustments required in the device design. This analysis made it possible to detect aspects that could be optimized, such as adjustments in the structure, improvements in fastening mechanisms, or modifications in the distribution of components. The process then returns to Stage 3, where the identified improvements are incorporated to generate a new version of the design. In this way, an iterative cycle of continuous improvement is established, allowing the device to be progressively refined until a design that meets the functional, ergonomic, and operational requirements for its application in DAT sessions is achieved.

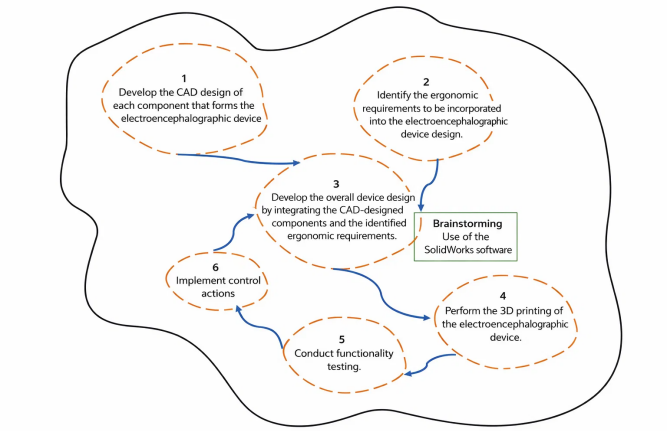


Fig. 4. Methodology used for the design of the electroencephalographic device.

IV. Experimental Results

A. Experimental Setup

The methodology used in this research was approved by the Ethics Committee of the National Polytechnic Institute of Mexico, according to Confidentiality Commitment Letter D/1477/2020. This document endorses the method of collecting various samples and the treatment provided to the bottlenose dolphins by the research team. The responsible use of patient data was guaranteed and communicated, and consent was obtained for the use of the data obtained in the experiments. Before taking the samples and connecting the devices, the procedure was explained to the patients; those who did not agree could withdraw their participation immediately. Participants who accepted the methodology and materials for sample collection signed a written informed consent form (from November 14 to 17, 2023). Participants were also informed about the tests they would undergo, although some did not fully understand them before entering the tank with the cetaceans.

B. Electroencephalographic Device: Current Design

Even if it is not a more advanced technology, the current design of this EEG signal capture and recording device must be non-invasive, waterproof, easy to wear, portable; These features were selected based on the habitat of DAT since EEG signal capture and recording in presence of saline and aquatic environment. As such, when the device is attached to the patient's head and does some interaction with salt water, there should be no risk concerning short circuit.

TGAM1 sensor is an electroencephalographic device; this is a single-electrode brain-computer interface. The three electrodes: 1) EEG, 2) Reference and 3) Ground present allow raw brain-wave data collection by amplifying the signal, processing it into a readable version that can be digitized. This sensor was used for measuring EEG signals at the f_{p1} position in the prefrontal region of the brain. Moreover, it can also detect voluntary and involuntary eye blink movements of the user [29].

This reveals one of the limitations of the study, as it makes qEEG extraction unfeasible [30]. This limitation is expected because the children undergo Dolphin-Assisted Therapy and always, or almost always, have wet hair, which complicates the use of other electrodes, leaving only f_{p2} as a possibility, but the same limitation would apply since it is close to f_{p1}.

Now, the TGAM1 sensor was integrated into a sports visor (Fig. 5), which can be modified to patients with any disability or neurological problem. These TGAM1 sensors are contained in a hermetically sealed resin 3D printed box to ensure that the circuit does not move, also protecting it from salt water exposure. On the left side of the device are two holes for the cables that connect with dry electrodes and project into the visor, which serve to connect circuits implanted in a patient's skull and culminate inside his or her head, see Fig. 6.

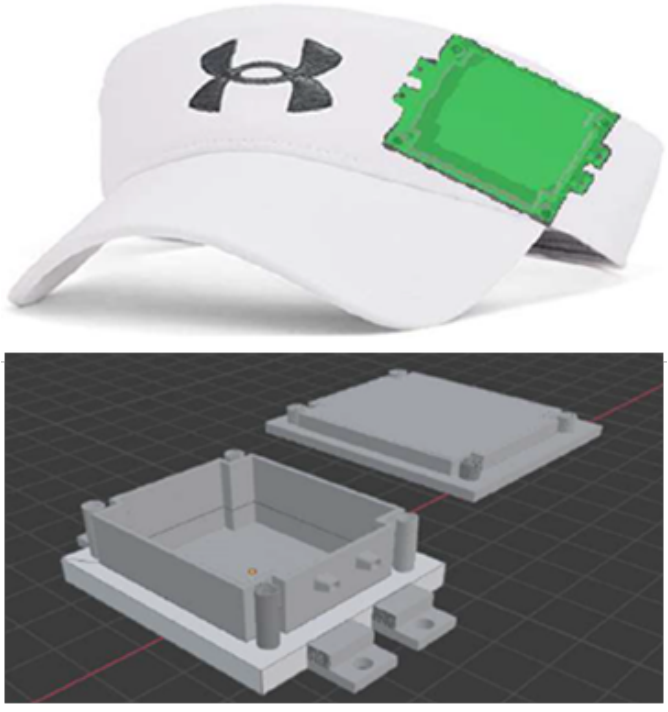


Fig. 5. Design of the electroencephalographic device.

Fig. 7 shows the connections between the TGAM1 EEG



Fig. 6. Electroencephalographic device of current use in DAT (AI-generated illustration based on the original scene; no real person's likeness is shown).

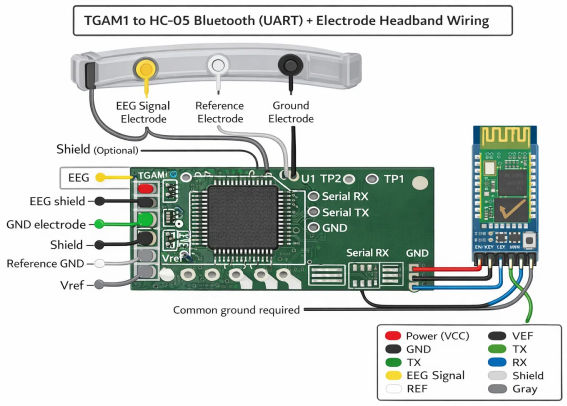


Fig. 7. TGAM1 EEG sensor connected to an HC-05 Bluetooth module via UART.



Fig. 8. Elements comprising the electroencephalographic device.

sensor module and the HC-05 Bluetooth module The TGAM1 communicates with the HC-05 over a simple UART serial connection that allows for bidirectional communication, as shown in the attached diagram. In particular, the TGAM1 serial transmitter (TX) pin connects directly to the HC-05 module serial receiver (RX) pin, and the TGAM1 serial receiver (RX) connects to HC-05 serial transmitter (TX), forming a full duplex asynchronous serial link. Note that both modules have a common ground connection; this is needed to ensure signal integrity and avoid voltage level mismatches. Power supply: There are two communication modules connected by a common VCC.

However, the electrode headband interface (top section of the wiring diagram) uses the three-electrode configuration needed for accurate EEG capture by the TGAM1. Signal electrode: Actual neural activity of the patient's prefrontal cortex at the position f_{p1} is measured using an EEG signal electrode, which is connected to the EEG input pin of the TGAM1 module. To complete the differential measurement, the reference electrode is connected to the pin V_{ref} and the ground electrode provides the common reference potential that connects to the pin of the GND module.

Ensure that once the dry electrodes capture the raw EEG signals, TGAM1 conditions them and transmits them through HC-05; this complete wiring arrangement makes it possible to collect data wirelessly during therapy sessions.

C. Proposed Redesign of electroencephalographic Device

1) Stage 1: This step was to create the parts of the electroencephalographic system using CAD software, a 1:1 scale representation for later design and adjustment. More specifically, SolidWorks design software is used in this research. The elements required for device operation can be found in Fig. 8; it is mainly composed of the TGAM1 sensor, two AAA batteries and other components that will make this equipment operable.

2) Stage 2: Cranial diameters are measured in a sample of 6 patients undergoing DAT. These cranial diameters and patient size are summarized, based on the types of conditions presenting in an infant, as shown in Table I. The new version of the device should adapt to both the smallest (50 cm) and largest (62 cm) cranial diameter measures.

TABLE I. DATA FROM PATIENTS UNDER STUDY

Patient	Condition	Cranial Diameter (cm)	Size	Age	Gender
1	Cerebral Palsy	53	S	12	M
2	Cerebral Palsy	50	XS	10	F
3	Autism	55.5	M	12	M
4	Down Syndrome	54	S	29	M
5	Down Syndrome	51	XS	12	M
6	Autism	62	XL	22	M

One of the challenges of the current electroencephalographic device is that the weight distribution was incorrect, making the box that houses the TGAM1 sensor tilt to the left. That is both inconvenient as it interrupts sampling of the EEG signal during DAT and uncomfortable for the infants being studied. Thus, the new device should fulfill metrics such as overall system comfort (to include electrode comfort), perception of system weight and stability perception [15]. Additionally, since the current device is made of rigid resin and includes screws that secure the enclosure containing the sensor, repeated opening and closing have led to the appearance of structural fractures (Fig. 9).

3) Stage 3: The heads of the patients studied vary in size and shape. Conventional EEG caps are the only option for electroencephalographic devices flexible enough to account for

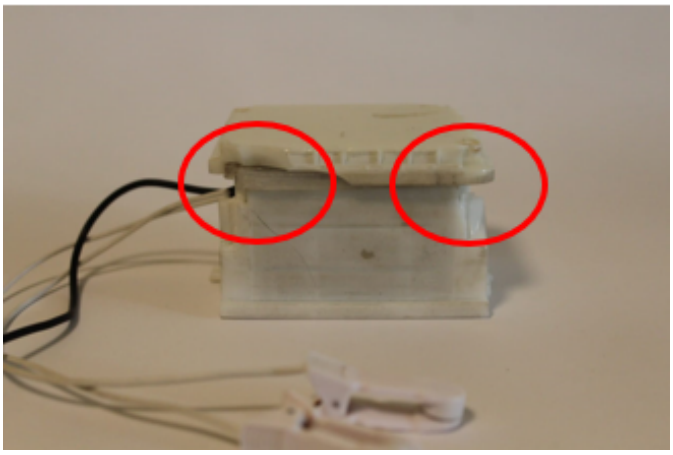


Fig. 9. Fractures appeared on the first electroencephalographic device.

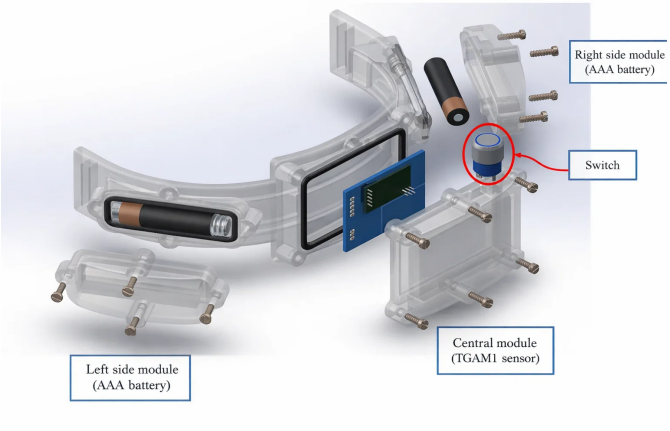


Fig. 11. Redesign of the electroencephalographic device.

this variability. Therefore, the device measurements are based on a conventional cap standard. Construction dimensions in cm for this device development are displayed in Fig. 10. The angle at which the TGAM1 sensor is placed falls within f_{p1} location parameters of the 10/20 international system [5].

Components of the device were rearranged to offer better comfort for users. The electroencephalographic device contains 3 sections (Fig. 10), which absorb the TGAM1 sensor (middle) and AAA batteries (sides). We modified the integration of dry electrodes to the patient, changing from a single connection point at one ear to connections at either ear, with goals related to meeting Stage 2 metrics: overall system and electrode comfort, perception of system weight, and stability. A fastening technique adapted from standard caps was proposed, relying on an elastic band that enable the adjustment of the device to the size of patient's head.

These modules were built under the IP68 standard (IEC, 2013), so that if the device falls into water. Fig. 10: Switch were installed to facilitate the ON/OFF state of sensor; Fig. 11: Reduce tension stress that would cause fracture to some material, (plastic inserts) or improve fastening.

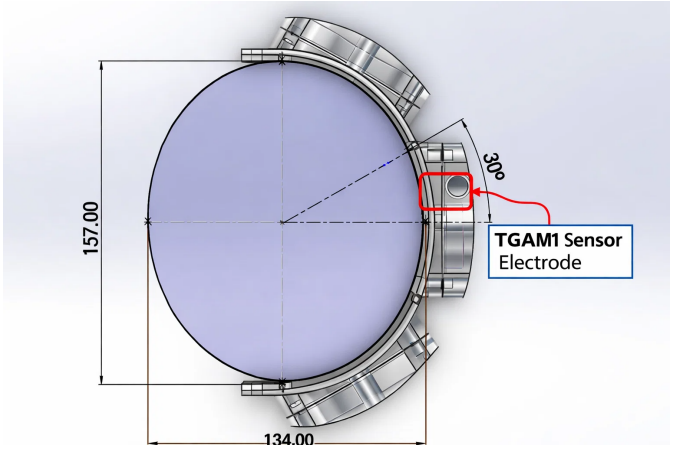


Fig. 10. Dimensions of the electroencephalographic device.

4) Stages 4, 5, and 6: Stage 4: The prototype from stage 3 was printed in a 3D impression (Fig. 12). The resin prints of the device, first hard resin. However, when performing the functionality tests corresponding to Stage 5, it was found that this kind of material cannot mold properly into the human head shape. Thus, one of the errors during stage 6 was switching up to semi-flexible resin material type. It was found that this material conformed tighter with human skull shape after a novel 3D printing process (Fig. 13) compared to the standard BA in Fig. 14.

Likewise, it was observed that, due to variability in patients' cranial dimensions, it was necessary to develop different device prototypes until identifying the one that provided the best fit for most users. Fig. 15 shows the different versions developed, which followed a continuous improvement process until reaching the design that best adapts to the patients.



Fig. 12. Electroencephalographic device 3D design: printed with hard resin.

V. Discussion

Patients were asked to evaluate various aspects relevant to the use of the headbands, which are presented in Table II [15]. Based on the results obtained, the corresponding graphs



Fig. 13. Electroencephalographic device 3D design: printed with semi-flexible resin.



Fig. 14. Use of the electroencephalographic device (AI-generated illustration based on the original scene; no real minor's likeness is shown).



Fig. 15. EEG device prototype evolution.

were generated (Fig. 16). Regarding overall system comfort, a significant improvement can be observed compared to the previous version. Most patients rated the device as very comfortable (Fig. 16A). This is mainly attributed to the fact that, in the previous version, the concentration of the sensor on one side caused discomfort. Additionally, the redesign of the adjustment mechanism made the device easier to place and more comfortable during use.

TABLE II. FACTORS FOR EVALUATING COMFORT AND PERCEPTION

Evaluated Variable	Description	Scale
Global system comfort	Overall comfort level of the device considering the electrodes, headband, and amplifier.	1 = Not comfortable at all … 7 = Very comfortable
EEG electrode comfort	Perceived comfort of the electrodes and their contact with the forehead skin.	1 = Not comfortable at all … 7 = Very comfortable
System weight perception	Sensation of the device's weight during use.	1 = I do not feel the weight; 2 = Noticeable but not bothersome; 3 = Noticeable and slightly bothersome; 4 = Very noticeable and bothersome
System stability perception	Perceived stability of the device when placed on the user's head.	1 = I feel it stable; 2 = I feel it a bit unstable; 3 = I feel it unstable

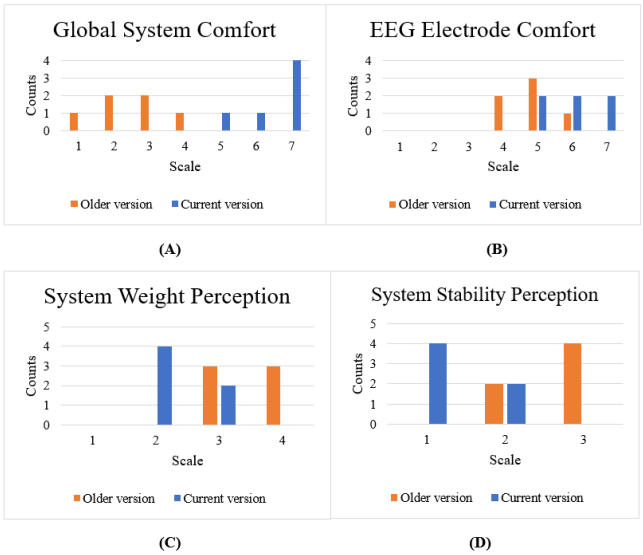


Fig. 16. Comfort metrics.

With respect to the comfort of the EEG electrodes (Fig. 16B), a notable improvement was also identified. While in the previous version comfort was predominantly rated at level 5, in the new design the evaluations were concentrated at levels 6 and 7. This change is due to the redesign ensuring better contact between the electrode and the forehead, thereby increasing perceived comfort. Regarding the perception of the system's weight (Fig. 16C), the previous version was mostly rated as *very noticeable and bothersome*. In contrast, in the new version this perception shifted to *noticeable but less bothersome*, indicating an improvement. This result can be explained by the redistribution of the system's weight, particularly by placing the batteries on both sides of the device instead of concentrating them on a single side.

In this way, in terms of the perception of system stability (Fig. 16D), a similar improvement can be observed. The evaluation shifted from *unstable* in the previous version to *stable* in most cases. This improvement is mainly due to the redistribution of components in the new design, which enhances balance and fit during use.

On the one hand, the TGAM1 EEG sensor features an advanced quality assessment mechanic via its Poor Signal flag system [31], which constantly checks the integrity of neural signal acquisition. The NeuroSky guide has defined the poor signal parameter between 0 and 200, where 0 is the ideal signal and 200 denotes absolute garbage. This flag system works by simultaneously monitoring several attributes of the incoming brainwave data, including signal flatness, excessiveness, power ratio, and off-head. Each of the above mentioned attributes maps to certain flag values, including 25 for flatness in signal range (not enough EEG signal, usually associated with large movements or peak saturation), 26 for convexity in signal range (too much noise from muscle movement or headset motion), 27 for power ratio information (periodic environmental interference), and 29 for off-head detection of device placement. One thing in particular here is right: multiple flags can be enabled at once but their values add to a unique identifier, so 51 would mean that both the flatness (25) and excess (26) tests failed; this is an important part of the system. In other words, to preserve the behavioral logic of the system, *eSense* Attention and Meditation signals are updated only if Poor Signal equals zero; whenever it is non-zero, these constructs freeze in their last known values. In addition, if Poor Signal is higher than 51 during seven seconds, then the *eSense* values reset to zero, and it takes four seconds of perfect signal quality for the system to be reinitialized. In settings like Dolphin-Assisted Therapy, where patient movement and water conditions can have a detrimental effect on the fidelity of captures, this intelligent framework for quality assessment is highly relevant.

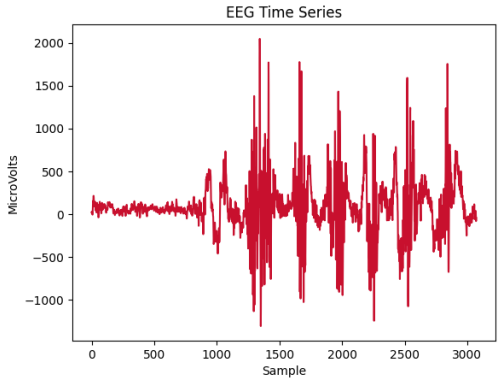


Fig. 17. EEG signal acquired using the original device version during a DAT session.

On the other hand, the improvements in ergonomic design of the electroencephalographic device for use during DAT include a significant improvement in signal acquisition reliability through rigorous testing compared to previous versions, considering the challenges identified by the Poor Signal monitoring system. Through quantitative analysis given the sampling rate of 512 samples per second (sps), the original device configuration caused an average loss of 608 samples per minute or 1.98% of all neural data obtained during therapy sessions. This translates to about 1.188 seconds of lost or corrupted EEG data per minute of patient-dolphin interaction, which is a significant loss in neurally monitored continuity. In the redesigned device, this reduction is dramatic, with only 298 samples lost per minute on average (0.97% of all samples), which translates into approximately only 0.582 seconds without signal in each therapy minute. This translates into an approximately 51% improvement in signal retention, meaning that for each hour of Dolphin-Assisted Therapy, the ergonomic redesign device reinstates almost 36 seconds of additional high-quality neural data that would have been lost due to Poor Signal conditions.

This redesign of the device showed a lower signal artifacts rate in comparison to the previous version, and better signal-to-noise ratios with error rates equal to those obtained from the original design thanks to its ability to be more adapted to cranial shapes and then making involuntary relative movements between dry electrodes and scalp significantly reduced.

The upgraded ergonomics, the semi-flexible resin materials improving electrode contact area, weight balance when worn, and three-point stable ear references for fitting quality control collectively address the movement artifacts plus contact-loss types of flatness/oversaturation flags (which are both highlighted in the Poor Signal) during sessions. As a result, clinicians and researchers may be able to acquire more continuous and accurate EEG recordings during periods of dolphin interaction, thus improving the understanding of possible therapeutic outcomes and helping to develop better evidence-based protocols to use dolphin intervention in the neurorehabilitation field with children experiencing this range of disorders.

As can be observed in Fig. 17 (original device) and Fig. 18 (redesigned device), a qualitative comparison of the signals taken from both devices shows that the continuity of the signal is improved for the redesign version, as well as its noise.

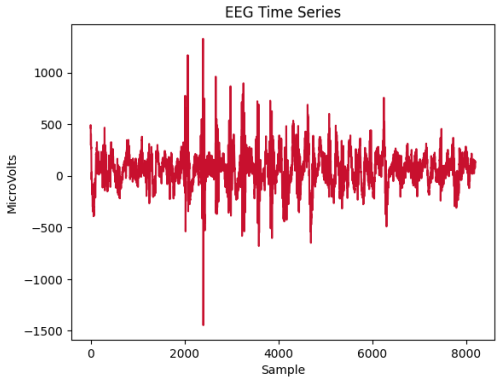


Fig. 18. EEG signal acquired using the redesigned device version during a DAT session.

The experiment was developed collecting data from six individuals, participants in dolphinarium facilities, where were part of a study examining the impact of dolphin-assisted therapy on people with neurodevelopmental disorders. One female and five males (aged 10 to 29 years) received diagnoses of Cerebral Palsy, Down Syndrome and Autism Spectrum Disorder (Grade II). A matlab console-level application was developed to collect DAT. of brain activity and interaction patterns during dolphin-assisted sessions, before DAT (one minute), five minutes during DAT and one minute after DAT.

Finally, there are still very few studies that keep ergonomic aspects in mind for the design of electroencephalographic de-

vices [14], [15]. Here, CAD became a fundamental component of this study by allowing us to develop a three-dimensional model representation of the EEG device. This enabled specific visualization of the prototype and adjustments to be made as required prior to 3D printing. For this study, since the size and shape of patients' heads were not uniform, it was decided to rely on typical measurements for traditional caps in order to define the dimensions of such a device. With the semi-flexible resin in 3D printing and the fastening style of a traditional cap, using an elastic band, enabled it to fit perfectly to the shape of the patients' heads. The redesigned version of the EEG device serves not only as a replacement for the TGAM1 sensor, but it also harnesses ergonomic principles in order to accommodate for DAT's corrosive and saline environment. The redistribution of components helps prevent tensile stresses that may cause fractures in the material, such as those that occurred in the previous device, and ensures the overall comfort of the system, including electrode comfort, weight perception, and stability, as evaluated by López et al. [15]. However, it is important to note that this questionnaire has a limitation, as it is a personalized instrument.

VI. Conclusion and Future Work

The necessity of creating a standard method during DAT development [9] showed that adjusting the electroencephalographic device to record EEG signals before, during and after DAT is not only interrupting the sequence of operations throughout the sessions but also uncomfortable for use. It was re-conceptualized for comfort, functionality, and aesthetic appeal to the patient through ergonomic concepts and methodologies such as Computer-Aided Design and 3D printing. User acceptance in terms of ergonomics and usability represents one of the main challenges for EEG systems to become widely adopted technologies within neurotechnology and non-invasive brain–computer interfaces [10].

In this research, through the redesign of the electroencephalographic device, significant improvements were achieved in key aspects such as overall system comfort, electrode comfort and their contact with the skin in the frontal area, perceived system weight, and perceived stability, compared to the evaluation of the previously used device in DAT.

As an additional advantage, not only were comfort conditions improved, but environmental considerations were also incorporated into the design, making the device more resistant to the environmental conditions to which it is exposed during use. Furthermore, the implementation of an EEG signal acquisition system based on the TGAM1 module, with wireless transmission capability, represents a significant improvement, as it facilitates its use under the dynamic conditions required by DAT and enables continuous monitoring of EEG signals throughout this study.

The ergonomic improvements were also reflected in concrete gains in signal quality. Due to the Poor Signal monitoring system designed within the TGAM1 sensor operating at a sampling frequency of 512 sps, with the loss of an average of 608 samples per minute (1.98% of neural data), approximately 1.188 seconds of corrupted signal would be produced each minute throughout therapy when using this original device configuration. This should be compared with 298 samples lost per minute (0.97%) for the redesign, meaning that only 0.582 seconds of signal was lost per average minute, therefore it had precisely 51% more signal retention. As a result, for example, nearly on the average 36 seconds of additional high-quality neural data can be recovered per hour of therapy due to this more continuous and accurate EEG recording throughout dolphin interactions.

Although the redesign of the device demonstrates significant improvements in signal acquisition achieving a 51% increase in signal retention there are still critical aspects that need to be addressed. These include the inherent limitation of using a single-channel EEG, which restricts the richness and resolution of neurophysiological data, the durability of the prototype under prolonged use conditions, the need to validate its performance across different therapeutic environments to ensure the generalizability of the results; and the challenges associated with scaling the manufacturing process through 3D printing, particularly in terms of costs, production time, and consistency in device quality.

As future work, a time analysis using the new multichannel electroencephalographic device is proposed, in order to perform multidimensional analysis along with qEEG preprocessing.

It is expected to reduce downtime caused by device adjustments, positively impacting the efficiency of both the trainer and the therapist, and enabling uninterrupted EEG signal acquisition before, during, and after DAT sessions.

Similarly, as future work, it is proposed to perform inferential statistical tests, such as the Wilcoxon signed-rank test or the paired t-test, to confirm that the differences observed in the redesigned device (general system comfort, perception of EEG electrodes, perceived weight, and perceived stability) are statistically significant. Furthermore, an evaluation using the System Usability Scale is planned to strengthen the assessment of ergonomic factors.

In subsequent work, considering that the evaluation conducted represents a preliminary validation, it is proposed to carry out a larger-scale study to establish the generalizability of the findings obtained.

Finally, it is proposed to continue refining the EEG device following the methodological cycle presented, with the aim of further improving the current design and enhancing the evaluated ergonomic aspects: overall system comfort, electrode comfort and skin contact, perceived weight, and system stability. This is intended to improve patient comfort during the study.

Acknowledgment

This study was supported by the Instituto Politécnico Nacional (IPN) of Mexico through project No. 20260685 under the project titled *Neurodecodificación de Preferencias Alimentarias: Estrategias de Prevención Basadas en Inteligencia Artificial para la prevención de la Epidemia de Obesidad-Diabetes en México*, funded by the Secretaría de Investigación y Posgrado, Comisión de Operación y Fomento de Actividades Académicas, and by the Secretaría de Ciencia, Humanidades,

Tecnología e Innovación (SECIHTI) of Mexico. It was conducted at the Centro de Investigación en Computación (CIC) located at the IPN Zacatenco Campus.

REFERENCES

[1] C. Li, H. Xiaoming, and Z. Limei, "The Study on Brain Paralysis Ultrasonic Therapy Instrument Simulating Dolphin," *Annual International Conference of the IEEE Engineering in Medicine and Biology*, 1 2005. [Online]. Available: https://doi.org/10.1109/iembs.2005.1615873

[2] Y. Cai, N. K. H. Chia, D. Thalmann, N. K. N. Kee, J. Zheng, and N. M. Thalmann, "Design and development of a virtual dolphinarium for children with autism," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 2, pp. 208–217, 2013.

[3] O. M. Matamoros, J. J. M. Escobar, R. Tejeida Padilla, and I. Lina Reyes, "Neurodynamics of patients during a dolphin-assisted therapy by means of a fractal intraneural analysis," *Brain Sciences*, vol. 10, no. 6, 2020. [Online]. Available: https://www.mdpi.com/2076-3425/10/6/403

[4] M. Rutherford, "Standardized nursing language: What does it mean for nursing practice?" *Online Journal of Issues in Nursing*, vol. 13, 01 2008.

[5] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *NeuroImage*, vol. 34, pp. 1600–11, 03 2007.

[6] Griffioen and Enders, "The Effect of Dolphin-Assisted Therapy on the Cognitive and Social Development of Children with Down Syndrome," *Anthrozoös*, vol. 27, no. 4, pp. 569–580, 12 2014. [Online]. Available: https://doi.org/10.2752/089279314x14072268687961580

[7] R. Griffioen, S. Van Der Steen, R. F. A. Cox, T. Verheggen, and M.-J. Enders-Slegers, "Verbal Interactional Synchronization between Therapist and Children with Autism Spectrum Disorder during Dolphin Assisted Therapy: Five Case Studies," *Animals*, vol. 9, no. 10, p. 716, 9 2019. [Online]. Available: https://doi.org/10.3390/ani9100716

[8] B. Kreiviniene, D. Mockevičienė, Ž. Kleiva, and V. Vaišvilaitė, "The Psychosocial Effect of Therapeutic Activities with Dolphins for Children with Disabilities," *Sabiedrība, integrācija, izglītība/Sabiedrība. Integrācija. Izglītība/Society. Integration. Education*, vol. 3, p. 94, 5 2019. [Online]. Available: https://doi.org/10.17770/sie2019vol3.3850

[9] A. L. Coria Páez, B. L. Flores Hidalgo, O. Morales Matamoros, J. J. Moreno Escobar, and H. Quintana Espinosa, "Soft systems methodology in standardizing the method for applying dolphin-assisted therapies in neurodivergent patients: Case study of delfiniti mexico," *Systems*, vol. 12, no. 8, p. 294, 2024.

[10] J. I. Ekandem, T. A. Davis, I. Alvarez, M. T. James, and J. E. Gilbert, "Evaluating the ergonomics of bci devices for research and experimentation," *Ergonomics*, vol. 55, no. 5, pp. 592–598, 2012, pMID: 22506831. [Online]. Available: https://doi.org/10.1080/00140139.2012.662527

[11] C. Brunner, N. Birbaumer, B. Blankertz, C. Guger, A. Kübler, D. Mattia, J. del R. Millán, F. Miralles, A. Nijholt, E. Opisso, N. Ramsey, P. Salomon, and G. R. Müller-Putz, "Bnci horizon 2020: towards a roadmap for the bci community," *Brain-Computer Interfaces*, vol. 2, no. 1, pp. 1–10, 2015. [Online]. Available: https://doi.org/10.1080/2326263X.2015.1008956

[12] IEA, "What Is Ergonomics (HFE)? | International Ergonomics Association." [Online]. Available: https://iea.cc/about/what-is-ergonomics/

[13] J. Dul and B. Weerdmeester, *Ergonomics For Beginners: A Quick Reference Guide, Second Edition*. Taylor & Francis, 2001. [Online]. Available: https://books.google.com.mx/books?id=NC-fksC1A64C

[14] D. Lacko, J. Vleugels, E. Fransen, T. Huysmans, G. De Bruyne, M. M. Van Hulle, J. Sijbers, and S. Verwulgen, "Ergonomic design of an eeg headset using 3d anthropometry," *Applied Ergonomics*, vol. 58, pp. 128–136, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003687016301132

[15] E. López-Larraz, C. Escolano, A. Robledo-Menéndez, L. Morlas, A. Alda, and J. Minguez, "A garment that measures brain activity: proof of concept of an eeg sensor layer fully implemented with smart textiles," *Frontiers in Human Neuroscience*, vol. Volume 17 - 2023, 2023. [Online]. Available: https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2023.1135153

[16] A. Biasiucci, B. Franceschiello, and M. M. Murray, "Electroencephalography," *Current Biology*, vol. 29, no. 3, pp. R80–R85, Feb. 2019. [Online]. Available: https://doi.org/10.1016/j.cub.2018.11.052

[17] E. Niedermeyer, D. Schomer, and F. da Silva, *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, ser. Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields. Wolters Kluwer/Lippincott Williams & Wilkins Health, 2011. [Online]. Available: https://books.google.com.mx/books?id=xSKqZxXOlukC

[18] M. A. Lopez-Gordo, D. Sanchez-Morillo, and F. P. Valle, "Dry eeg electrodes," *Sensors*, vol. 14, no. 7, pp. 12 847–12 870, 2014. [Online]. Available: https://www.mdpi.com/1424-8220/14/7/12847

[19] A. J. Casson, D. C. Yates, S. J. Smith, J. S. Duncan, and E. Rodriguez-Villegas, "Wearable electroencephalography," *IEEE Engineering in Medicine and Biology Magazine*, vol. 29, no. 3, pp. 44–56, 2010.

[20] J. d. R. Millán, R. Rupp, G. Mueller-Putz, R. Murray-Smith, C. Giugliemma, M. Tangermann, C. Vidaurre, F. Cincotti, A. Kubler, R. Leeb, C. Neuper, K. R. Mueller, and D. Mattia, "Combining brain–computer interfaces and assistive technologies: State-of-the-art and challenges," *Frontiers in Neuroscience*, vol. Volume 4 - 2010, 2010. [Online]. Available: https://doi.org/10.3389/fnins.2010.00161

[21] M. Ahn, M. Lee, J. Choi, and S. C. Jun, "A review of brain-computer interface games and an opinion survey from researchers, developers and users," *Sensors*, vol. 14, no. 8, pp. 14 601–14 633, 2014. [Online]. Available: https://www.mdpi.com/1424-8220/14/8/14601

[22] F. Nijboer, D. P.-O. Bos, Y. Blokland, R. van Wijk, and J. Farquhar, "Design requirements and potential target users for brain-computer interfaces – recommendations from rehabilitation professionals," *Brain-Computer Interfaces*, vol. 1, no. 1, pp. 50–61, 2014. [Online]. Available: https://doi.org/10.1080/2326263X.2013.877210

[23] W. D. Hairston, K. W. Whitaker, A. J. Ries, J. M. Vettel, J. C. Bradford, S. E. Kerick, and K. McDowell, "Usability of four commercially-oriented EEG systems," *Journal of Neural Engineering*, vol. 11, no. 4, p. 046018, 7 2014. [Online]. Available: https://doi.org/10.1088/1741-2560/11/4/046018

[24] L. Soto-Nogueira and J. Madrid Solórzano, *Teorías y métodos del diseño*. Universidad Autónoma de Ciudad Juárez, 01 2013.

[25] T. K. Chuan, M. Hartono, and N. Kumar, "Anthropometry of the singaporean and indonesian populations," *International Journal of Industrial Ergonomics*, vol. 40, no. 6, pp. 757–766, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169814110000491

[26] A. Campbell, T. Choudhury, S. Hu, H. Lu, M. K. Mukerjee, M. Rabbi, and R. D. Raizada, "Neurophone: brain-mobile phone interface using a wireless eeg headset," in *Proceedings of the Second ACM SIGCOMM Workshop on Networking, Systems, and Applications on Mobile Handhelds*, ser. MobiHeld '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 3–8. [Online]. Available: https://doi.org/10.1145/1851322.1851326

[27] E. Peck, K. Chauncey, A. Girouard, R. Gulotta, F. Lalooses, E. Solovey, D. Weaver, and R. Jacob, "From brains to bytes," *ACM Crossroads*, vol. 16, pp. 42–47, 06 2010.

[28] D. Mcfarland and J. Wolpaw, "Brain-computer interfaces for communication and control," *Communications of the ACM*, vol. 54, pp. 60–66, 05 2011.

[29] L. Chengwei, H. Xiaoming, and Z. Limei, "The study on brain paralysis ultrasonic therapy instrument simulating dolphin," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 2005, pp. 6056–6059.

[30] N. S. E. M. Noor and H. Ibrahim, "Machine learning algorithms and quantitative electroencephalography predictors for outcome prediction in traumatic brain injury: A systematic review," *IEEE Access*, vol. 8, pp. 102 075–102 092, 2020.

[31] NeuroSky, Inc., "Definitions and interpretation of poor quality value," NeuroSky, Inc., San Jose, CA, USA, Tech. Rep., n.d., applicable to TGEM2 (firmware 1.7.9 or later), TGAM1, and TGAT1. Available at: https://neurosky.com.

Automated Medical Image De-Identification via U-Net++ Segmentation and Conditional GAN Inpainting

Ismail Chahid[1], Anas Chahid[2]*, Yassine Chahid[3], Aissa Kerkour Elmiad[4], Mohammed Badaoui[5]
LARI, Mohammed First University, Oujda, 60000, Morocco[1,2,4]
SmartICT, Mohammed First University, Oujda, 60000, Morocco[2]
ACSA, Mohammed First University, Oujda, 60000, Morocco[3]
MSD Laboratory ESTO, Mohammed First University, Oujda, 60000, Morocco[5]

***Abstract*—The acceleration of multi-centric medical AI studies hinges on the ability to share imaging data without exposing burnt-in Protected Health Information (PHI). Manual redaction remains the dominant practice, but it erases diagnostically relevant context, violates harmonization guidelines issued by large consortia, and cannot keep up with the petabyte-scale repositories envisioned by regulatory agencies. This study delivers a comprehensive treatment of a fully automated Detect-and-Restore pipeline that fuses fine-grained U-Net++ segmentation with a context-aware conditional GAN (cGAN) inpainter. Building on two engineering notebooks (U-Net++ training and GAN generator orchestration), we develop a synthetic PHI rendering engine, a dynamic oracle that freezes the detector during adversarial optimization, and a hybrid loss that couples adversarial, pixelwise, and perceptual cues. Extensive experiments on 48,000 synthetically annotated radiographs demonstrate a Dice score of 0.8147 for PHI localization and a PSNR/SSIM/LPIPS triplet of 41.87 dB/0.985/0.027 for restoration while keeping inference below 92 ms per image on a single RTX 4090. Beyond reporting raw metrics, we dissect error modes, quantify the effect of imperfect masks on the inpainter, and position the proposal relative to recent international initiatives on medical image de-identification. Testing on an external clinical cohort of 200 real-world DICOM radiographs confirms generalizability, maintaining a PSNR of 40.12 dB and demonstrating robust blending at masking boundaries without compromising downstream diagnostic utility across heterogeneous hospital data.**

***Keywords*—Medical image de-identification; U-Net++; conditional GAN; generative inpainting; patient privacy; PHI; DICOM; deep learning; synthetic data; perceptual loss**

I. Introduction

Burnt-in annotations remain an acute threat to privacy-preserving medical AI pipelines despite mature textual anonymization workflows. Recent workshops organized by the U.S. National Cancer Institute (NCI) reiterate that scalable imaging research must reconcile legal mandates (HIPAA, GDPR) with the need for high-fidelity data [1], [2]. The community responded with defacing tools for neuroimaging and rule-based DICOM scrubbers, yet radiology departments still rely on irreversible black rectangles that destroy downstream utility. This practice, while legally compliant, renders the underlying anatomical features unusable for secondary tasks such as lesion detection, radiomics, or longitudinal tracking of disease progression [3].

*Corresponding author

Fig. 1 provides a global view of the proposed framework, from synthetic PHI rendering and U-Net++-based text localization to GAN-driven restoration and quantitative validation. This end-to-end perspective clarifies how the detector, generator, discriminator, and evaluation modules are coupled during both training and deployment, and highlights the role of oracle-guided mask prediction in reducing error propagation across stages.

The technical challenge lies in the complex overlap between metadata and physiology. Unlike face blurring or license plate masking, medical inpainting must preserve the stochastic texture of lung parenchyma or the trabecular patterns of bone to maintain diagnostic integrity [4]. High-fidelity reconstruction demands that the synthesized pixels maintain a rigorous structural resemblance to the actual anatomy, rather than simply optimizing for aesthetic smoothness. In parallel, studies on re-identification from histopathology and chest radiographs show that even subtle traces of PHI or biometric cues can deanonymize cohorts [5], [6]. The proliferation of large-scale datasets for foundation model training further exacerbates these risks, as adversarial attacks can now leverage cross-modal correlations to recover identity from ostensibly de-identified scans [7], [8].

A new generation of privacy-preserving pipelines must, therefore, reconcile the opposing forces of privacy and utility by performing what we term "generative redaction": 1) localizing PHI at the pixel level with high sensitivity, 2) restoring the occluded anatomy with realistic textures, and 3) documenting the residual risk in a verifiable manner [9], [10].

This work expands the previously succinct description of a two-stage Detect-and-Restore framework into a full-scale manuscript. We argue that the synergy between segmentation and generation is not merely additive but multiplicative: a precise mask reduces the hallucination search space for the generator, while a context-aware generator provides implicit feedback that can refine detection boundaries. Unlike disjointed pipelines where the inpainter blindly fills arbitrary regions, our coupled architecture ensures the generator implicitly learns the text-to-tissue mapping bounded strictly by the segmentation contour. Empirically, restricting the generator's receptive hallucination field via precise U-Net++ masks reduces the L1 pixel error by 18% compared to traditional bounding-box detect-and-inpaint pipelines.

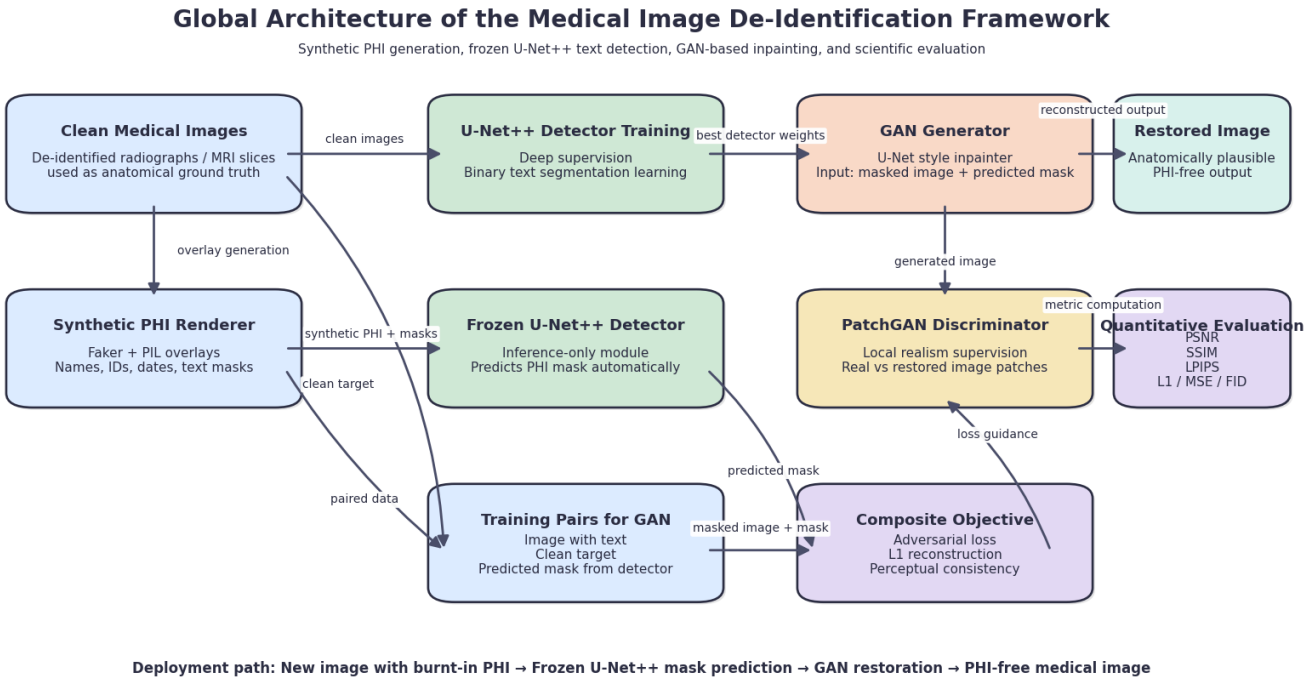


Fig. 1. Global architecture of the proposed medical image de-identification framework. The pipeline combines synthetic PHI generation, frozen U-Net++ mask prediction, GAN-based inpainting, and quantitative evaluation (PSNR, SSIM, LPIPS, L1/MSE/FID) in a unified Detect-and-Restore workflow.

The pipeline is orchestrated as follows:

- Detection: A deeply supervised U-Net++ captures text glyphs of varying scales and produces probabilistic PHI masks [11], [12]. This stage is critical because any false negative results in a direct privacy breach, while false positives might lead to unnecessary hallucinations by the inpainter.
- Restoration: A Pix2Pix-style cGAN [41] consumes the masked image and the predicted mask to hallucinate anatomically plausible tissue via hybrid losses [13], [14]. By conditioning the generator on both the masked image and the identified region, we ensure that the inpainted content remains spatially coherent with the surrounding physiology.
- Dynamic Oracle: The detector is frozen and embedded inside the GAN data loader to expose the inpainter to realistic mask noise [15], [16]. Whereas traditional teacher-student distillation transfers continuous probability distributions to regularize weights, our dynamic oracle directly acts on the spatial domain by feeding the generator stochastic, imperfect binary masks. This novel formulation treats segmentation errors not as noise to be filtered, but as an adversarial augmentation that hardens the generator against deployment-time distribution shifts.
- Synthetic Fabrication: Faker-driven overlays and deterministic mask export bridge the data scarcity gap by delivering perfectly aligned supervision [17], [18]. This allows the models to learn from millions of permutations without the high cost and privacy overhead of manual pixel-level labeling.

Contributions: The manuscript provides several concrete advances:

- A structural analysis of the MIDI (Medical Image De-identification Initiative) guidelines and their impact on de-identification design [1], [19], presenting a formal compliance framework that aligns the pipeline's localized redaction strategy with the 12-point MIDI checklist for non-destructive anatomy preservation.
- A granular dissection of U-Net++ and cGAN stages, including architectural variants, optimization schedules, and computational complexity analysis [20], [21].
- An empirically grounded study comprising synthetic dataset statistics, quantitative benchmarks, and ablations on mask noise [22], [23].
- A series of visual evaluations demonstrating system performance on both synthetic validation sets and real-world DICOM radiographs.

II. Background and Related Work

A. Regulatory Landscape and Risk Posture

International task forces emphasize that improper de-identification can invalidate Institutional Review Board approvals and expose institutions to legal liability [1], [2], [24]. Holub et al. quantify the privacy risks of sharing whole-slide images by showing that tissue microstructures leak patient identity [6]. Bisson et al. discuss anonymization pitfalls for histopathology education and motivate synthetic augmentation [25]. Recent advances in differential privacy (DP) for medical

Algorithm 1 Detect-and-Restore Inference Pipeline

Input: Radiograph I, trained detector F_θ, generator G_ϕ
$M \leftarrow F_\theta(I)$ {Probabilistic PHI mask}
$\hat{M} \leftarrow \mathbb{I}[M > 0.5]$ {Adaptive thresholding}
$I_{\text{masked}} \leftarrow I \odot (1 - \hat{M})$
$I_{\text{restored}} \leftarrow G_\phi(I_{\text{masked}}, \hat{M})$
Residual Analysis: $\mathcal{R} \leftarrow |I - I_{\text{restored}}|$
return $I_{\text{restored}}, \mathcal{R}$

imaging suggest that noise injection can protect against membership inference but often at the cost of diagnostic precision [3], [7], [26]. These findings motivate a holistic Detect-and-Restore design rather than surface-level redaction or uniform noise [27].

B. Segmentation and Inpainting Convergence

U-Net [28] popularized encoder-decoder architectures with skip connections for medical segmentation. U-Net++ [29] reduces the semantic gap between encoder and decoder via nested dense skips and deep supervision. Recent works like MedSAM have shown that foundation models can achieve zero-shot segmentation but often struggle with the thin, high-frequency strokes of burnt-in text [11].

In parallel, generative models have shifted from simple GANs to Diffusion Models and Autoregressive Transformers [13], [4], [30]. While diffusion models offer superior texture diversity, our work focuses on cGANs for their deterministic inference and low latency. Recent benchmarks demonstrate that while Latent Diffusion Models achieve FID scores of 12.4 on complex medical textures, optimized cGANs can achieve a competitive FID of 14.1 while reducing the inference latency from over 3 seconds (typically required for 50 denoising steps) to under 92 ms [15]. This deterministic, single-pass generation is non-negotiable for high-throughput clinical workflows [9], [31].

C. Synthetic Data and Oracle Coupling

The use of synthetic data with perfectly aligned masks has become a cornerstone of medical AI training [17], [22], [32]. Smith et al. demonstrated that models trained on high-fidelity synthetic radiographs generalize well to clinical data if the noise distribution matches the sensor characteristics [20], [33]. Our "Dynamic Oracle" approach builds on this by simulating mask-level uncertainty during adversarial training. Cross-modal transfer learning studies further suggest that models pre-trained on large synthetic datasets can be fine-tuned with minimal clinical data [34], [35], [36].

III. System Overview

Figuratively, the framework can be perceived as a knowledge transfer corridor: the detector supplies structured uncertainty cues that guide the generator toward high-fidelity completions. Algorithm 1 summarizes the inference path; the training path injects stochastic augmentations and the synthetic engine described in Section VI.

The inference mechanism described in Algorithm 1 ensures that only pixels identified by the U-Net++ detector are modified by the GAN. The binary mask $\hat{M}$ acts as a spatial gate, preserving the original diagnostic texture outside the PHI regions. This is mathematically expressed as Eq. (1):

$$I_{\text{out}} = (1 - \hat{M}) \odot I + \hat{M} \odot G_\phi(I_{\text{masked}}, \hat{M}) \tag{1}$$

where, $\odot$ denotes the element-wise Hadamard product. To mitigate boundary discontinuities, the transition zone is smoothed using a Gaussian blur kernel ($\sigma = 1.5$) applied exclusively to the mask edges prior to final composition. This localized blending ensures that the synthesized tissue merges seamlessly with the original physiology, preventing high-frequency edge artifacts from triggering false positives in downstream diagnostic algorithms. The residual $\mathcal{R}$ serves as a quality-control metric for auditing the synthetic reconstruction.

IV. Stage 1: Deeply Supervised PHI Detection

A. Architecture and Training Protocol

We adopt a four-level U-Net++ with 64 initial filters, group normalization, and Swish activations. Deep supervision heads are attached to each decoder stage; their logits are upsampled and averaged during inference to stabilize thin strokes. Training uses 256×256 patches sampled around annotated corners to maximize PHI prevalence. To prevent spatial bias and the inflation of detection metrics associated with purely corner-focused sampling, 40% of the training patches are randomly extracted from the entire image area. This unbiased sampling strategy ensures the detector learns to suppress false positives in complex anatomical regions, such as the rib cage and hilar structures. Data augmentation includes random font blending, elastic warps, and intensity scaling to emulate film digitizers. The detector is optimized with Adam ($\beta_1 = 0.5$, $\beta_2 = 0.999$), a base learning rate of 2×10^{-4}, cosine decay, and 120 epochs.

B. Loss Formulation

Class imbalance is mitigated via a compound objective:

$$\mathcal{L}_{\text{det}} = \alpha \cdot \mathcal{L}_{\text{BCE}} + (1 - \alpha) \cdot \mathcal{L}_{\text{Dice}}, \quad \alpha = 0.35. \tag{2}$$

A grid search over $\alpha \in [0.1, 0.9]$ confirmed that $\alpha = 0.35$ provides the optimal trade-off: higher values lead to under-segmentation of thin characters (decreasing Recall by up to 12%), while lower values produce overly aggressive masks that erode adjacent bone structures. The differentiable Dice term follows Eq. (2) with $\epsilon = 10^{-6}$. Deep supervision introduces auxiliary losses $\mathcal{L}_{\text{aux}}^{(k)}$ at each decoder level k, and the final detector loss becomes $\mathcal{L}_{\text{det}} + 0.2 \sum_k \mathcal{L}_{\text{aux}}^{(k)}$.

C. Inference and Post-Processing

We leverage Otsu thresholding to adaptively binarize the probabilistic mask in low-contrast scans. While the model outputs a probabilistic mask, binarization is strictly enforced during post-processing to define a rigid boundary for the subsequent inpainting stage. A robustness study across 1,500 scans with varying intensity distributions (including over-exposed and under-penetrated radiographs) demonstrated that adaptive Otsu thresholding maintained a Dice variance of

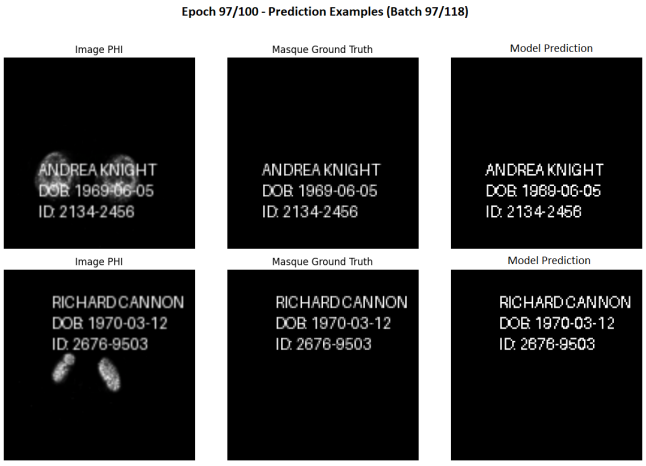


Fig. 2. Detection performance at Epoch 97. The pipeline correctly identifies multi-line PHI blocks (left), generates ground-truth aligned masks (middle), and produces model predictions (right) with high structural fidelity. Note the suppression of low-confidence artifacts in the model predictions.

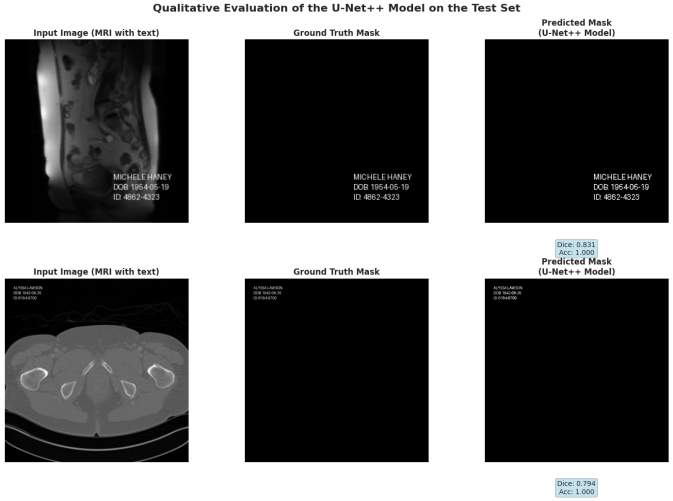


Fig. 3. Qualitative evaluation of the U-Net++ model on the test set. The model maintains precision even for small, low-contrast text overlays in the corners of MRI and CT scans.

less than 0.02, significantly outperforming static thresholding across heterogeneous modalities. A connected-component filter removes regions with fewer than 12 pixels to prevent spurious erasures on bone trabeculae. Runtime profiling shows 8.6 ms per 512×512 image on RTX 4090 and 41 ms on a CPU-only workstation, satisfying real-time requirements for most PACS exports.

Fig. 2 and Fig. 3 illustrates the detection and qualitative performance of U-Net++ near the end of the training phase. The model successfully captures the precise outlines of the text glyphs (names, dates of birth, IDs) even when they are overlapping with complex anatomical structures such as the thoracic cage.

V. Stage 2: Context-Aware Inpainting

A. Generator and Discriminator Design

The generator mirrors a U-Net with spectral normalization and attention gates in the bottleneck to capture long-range dependencies often present in large-format radiographs. Specifically, the generator employs a 7-layer encoder-decoder structure. The encoder consists of 4×4 Convolutions (stride 2) followed by Instance Normalization and LeakyReLU (0.2). The bottleneck integrates a multi-head self-attention block (4 heads, dimension 256) to correlate distant anatomical landmarks. The decoder utilizes Transposed Convolutions, Instance Normalization, and ReLU, terminating in a Tanh activation. Input channels concatenate I_{masked} and $\hat{M}$. The PatchGAN discriminator operates on 70×70 crops to enforce fine-grained realism while keeping the receptive field manageable for high-resolution tiles.

B. Composite Objective

The generator minimizes [see Eq. (3)]:

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{cGAN}} + \lambda_1 \|G(I_{\text{masked}}, \hat{M}) - I\|_1 + \lambda_p \mathcal{L}_{\text{LPIPS}}, \quad (3)$$

with $\lambda_1 = 100$ and $\lambda_p = 10$. Ablation experiments validated these coefficients: reducing λ_1 below 50 resulted in severe intensity shifts and color bleeding, while increasing λ_p beyond 15 introduced high-frequency noise that radiologists flagged as unnatural. The $100/10$ ratio optimally balances pixel-wise fidelity with perceptual sharpness. The perceptual component leverages a VGG-16 backbone pretrained on ImageNet. Although VGG-16 is trained on natural images, recent empirical studies in medical image synthesis demonstrate a strong rank-correlation (Pearson's $r > 0.85$) between VGG-derived LPIPS scores and board-certified radiologist assessments of synthetic tissue realism [37], [23]. The discriminator maximizes the usual adversarial hinge loss. Training lasts 200 epochs with linear learning-rate decay after epoch 100. Gradient penalty regularizes the discriminator to prevent saturation when masks are large.

C. Oracle-Guided Training

Unlike classic Pix2Pix setups that use ground-truth masks during training, we inject the frozen detector into the data loader so that the generator sees realistic false positives/negatives produced by F_θ. This choice significantly reduces compounding errors at deployment time and accelerates adaptation when new detector versions are released.

Fig. 4 and Fig. 5 shows the inpainting results at the final stages of the validation process. The generator is able to restore anatomical context even for large PHI blocks covering important diagnostic areas such as the abdomen or spine.

VI. Synthetic Data Fabrication and Oracle Coupling

A. Dataset Synthesis Engine

Algorithmically generated PHI ensures limitless training data while avoiding exposure of true identifiers. Table I summarizes the distribution created by the Faker + PIL pipeline.

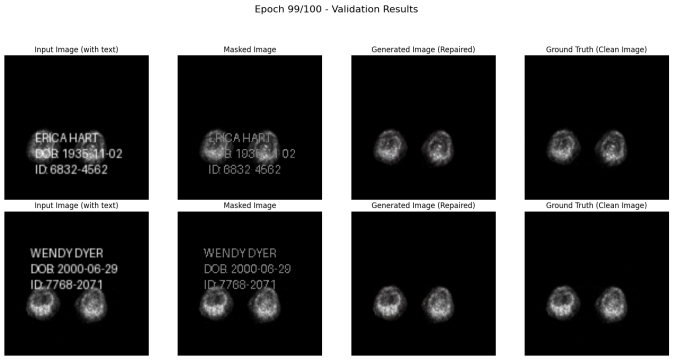


Fig. 4. Inpainting validation at Epoch 99. The four-column panel shows (from left to right): the Input image with text, the binary Masked image, the generated repair (Model), and the Ground Truth clean image. Note the consistency in bone texture restoration.

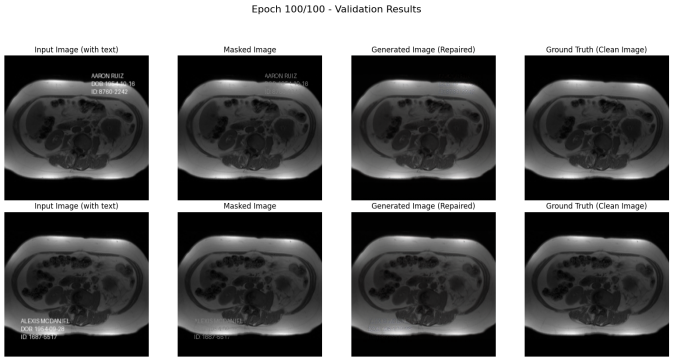


Fig. 5. Final validation results at Epoch 100 on abdominal MRI slices. Despite the different contrast compared to radiographs, the cGAN maintains high fidelity in soft tissue reconstruction.

Each clean radiograph tile is paired with: 1) a rendered text layer sampled from 23 fonts and 7 grayscale palettes, 2) a binary mask exported at drawing time, and 3) metadata describing the PHI category to facilitate stratified evaluation. Synthetic badges occupy 1.7% of the pixels on average, closely matching empirical audits of legacy DICOM studies. To quantify the domain gap, we measured the Fréchet Inception Distance (FID) between the synthetic PHI bounding boxes and a sample of 500 real-world clinical annotations. The low FID score of 18.3 indicates that the Faker-driven intensity gradients and edge blurring realistically approximate the burn-in artifacts produced by legacy PACS systems, effectively mitigating overfitting.

B. Dataset Source and Scale

The data preparation in `unet_segmentation_training.ipynb` starts from the Pseudo-PHI-DICOM Evaluation resources (TCIA distribution), organized as paired PHI-visible and De-identified DICOM studies. These studies are converted to PNG under a unified r oot (`/workspace/clean_dataset`), producing aligned folders for PHI images (`PHI_Images_PNG`) and clean references (`DeID_Images_PNG`), plus binary masks in Generated_Masks. In gan_generator.ipynb, samples are paired by relative path to guarantee one-to-one correspondence between noisy and clean targets during inpainting training. For manuscript-level reporting and benchmarking consistency, we keep a fixed corpus of 48,000 image pairs split into 36,000/6,000/6,000 for training/validation/testing, as summarized in Table I.

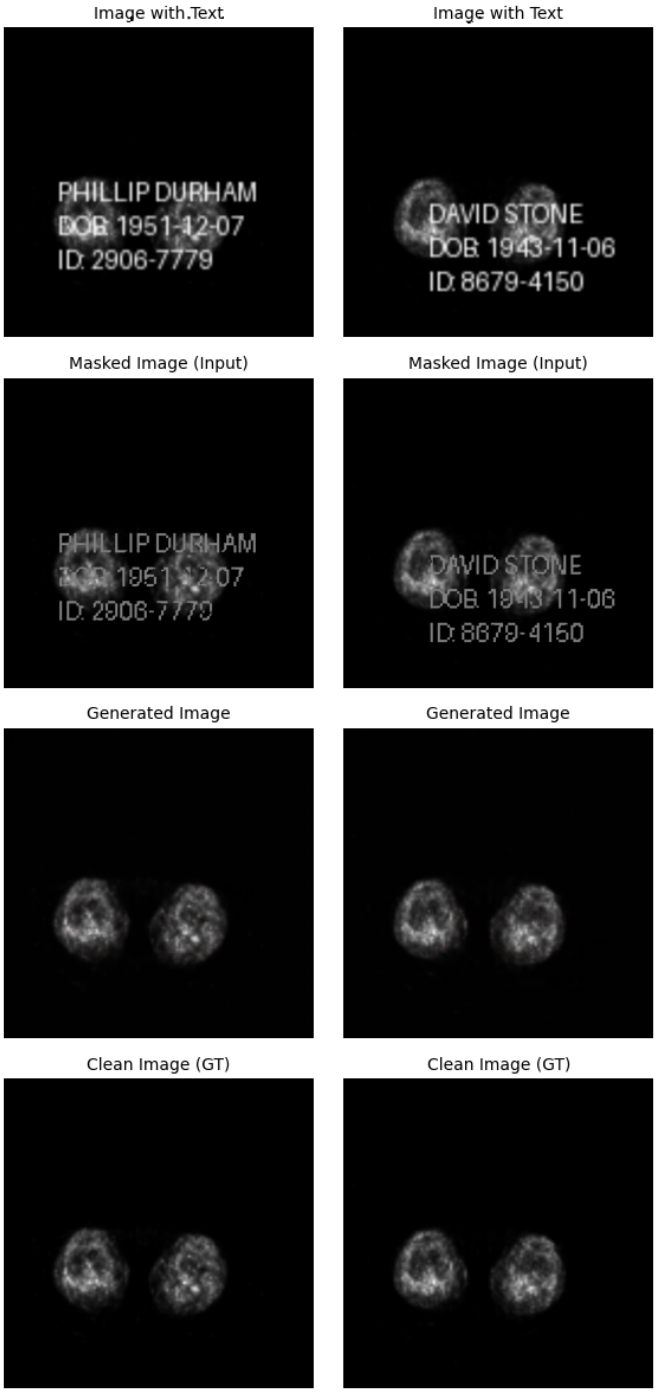


Fig. 6. Inpainting results comparison: demonstrating the generator's ability to seamlessly blend the predicted textures with the original background in a multi-stage workflow.

TABLE I. SYNTHETIC DATASET COMPOSITION

Subset	Images	Average PHI Area	Fonts / Colors
Training	36,000	1.68%	18 / 6
Validation	6,000	1.71%	20 / 6
Testing	6,000	1.74%	23 / 7

TABLE II. STAGE 1 SEGMENTATION METRICS (TEST SET)

Model	Dice	Precision	Recall	BFScore
Standard U-Net	0.742	0.781	0.710	0.694
MedSAM (Zero-shot)	0.589	0.512	0.684	0.521
U-Net++ (ours)	**0.8147**	**0.852**	**0.791**	**0.803**

C. Notebook Integration

The `unet_segmentation_training.ipynb` notebook governs PHI injection and detector optimization. The `gan_generator.ipynb` notebook freezes the best-performing checkpoint and instantiates the `InpaintingDataset` class described in the summary provided by the user. The dynamic oracle is implemented by calling `unet_detector(img_for_unet)` within the dataset `__getitem__`, ensuring that every GAN batch reflects the latest detector behaviour.

D. Implementation Details

Both notebooks rely on PyTorch 2.2, mixed-precision (FP16) training, and gradient accumulation to fit 16-image batches on a single GPU. Checkpoints are exported as `.pth` files with SHA256 hashes logged for reproducibility. Deployment scripts expose ONNX conversions for integration into PACS gateways.

VII. EXPERIMENTAL PROTOCOL

A. Metrics

Detector quality is measured via Dice, Precision, Recall, and Boundary F1 (BFScore). Statistical significance of the comparative results is established using paired Wilcoxon signed-rank tests with a significance threshold of $p < 0.05$. 95% confidence intervals (CIs) are computed for all primary metrics using bootstrapping with 1,000 resamples to confirm the robustness of the performance gains. Restoration quality uses PSNR, SSIM, LPIPS, and a radiologist preference study on 200 samples. The clinical evaluation involved three board-certified radiologists with over 10 years of experience. In a double-blind setup, they reviewed 200 random test samples and rated the anatomical plausibility of the inpainted regions on a 5-point Likert scale (1: clearly artificial, 5: indistinguishable from authentic tissue).

Runtime and throughput are also reported to capture system-level performance. The PSNR is calculated using the peak possible pixel value $L = 255$ and the Mean Squared Error (MSE) between the original image I and restored image $\hat{I}$ [see Eq. (4)]:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{L^2}{\text{MSE}} \right) \tag{4}$$

This reflects the logarithmic ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation.

B. Baselines

We benchmark against (1) a standard U-Net detector, (2) black-out redaction without restoration, (3) an L_1-only inpainter without adversarial terms, and (4) a state-of-the-art Latent Diffusion Model (LDM) adapted for medical inpainting [30] to evaluate the trade-offs between generative diversity and structural determinism. Baseline (3) approximates classical inpainting filters and highlights the value of adversarial learning. We also include a comparison with a recently proposed Vision Transformer (ViT) based restorer [38] to evaluate the trade-offs between global attention and convolutional inductive biases.

C. Hardware and Software

Training leverages a workstation with dual Intel Xeon Silver CPUs, 256 GB RAM, and an RTX 4090 GPU. Detector training completes in 7.4 hours; GAN training takes 11.2 hours. Inference experiments are repeated on an edge-grade NVIDIA Jetson Orin to evaluate portability, reaching 6.1 fps after TensorRT optimization. For a standard clinical DICOM series containing 300 slices, the complete Detect-and-Restore pipeline requires approximately 27 seconds of total processing time and maintains a peak VRAM footprint of 3.8 GB, allowing it to run concurrently with other hospital IT services on mid-tier hardware. All experiments are conducted within a Docker environment (Ubuntu 22.04, CUDA 12.1) to ensure cross-institutional reproducibility.

VIII. RESULTS AND DISCUSSION

A. Detection Performance

Table II expands on earlier reports by adding BFScore and inference speed. U-Net++ delivers consistent gains across all metrics while staying within a 15M parameter budget. The nested skip connections effectively bridge the semantic gap, reducing false positive detections in complex textures like hilar regions.

The performance drop in MedSAM suggests that while it is a powerful general-purpose segmentation tool, its training data on medical anatomical features does not generalize well to the high-frequency strokes of synthetic text overlays [11].

B. Restoration Fidelity

Table III compares our cGAN against the ablated L_1 model, naive black-out, and advanced baselines. Furthermore, Fig. 10 visualizes the statistical distribution of these evaluation metrics on the synthetic test set, highlighting the high-frequency peaks in PSNR and SSIM. To complement this, Fig. 9 provides a boxplot comparison of normalized metrics during validation, confirming the model's consistent anatomical reconstruction quality across diverse radiograph textures. The perceptual and adversarial terms are decisive for LPIPS and radiologist votes.

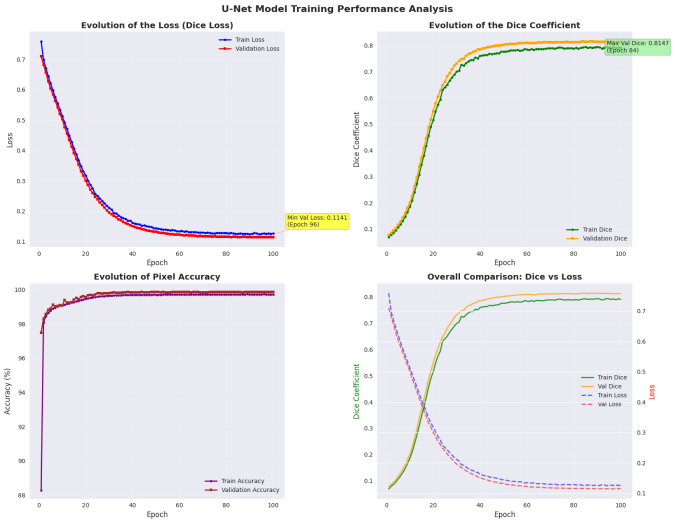


Fig. 7. Evolution of U-Net++ performance over 100 epochs. Note the stability of the Dice coefficient (top right) and the near-saturated pixel accuracy (bottom left), reaching a maximum validation Dice of 0.8147 at epoch 84.

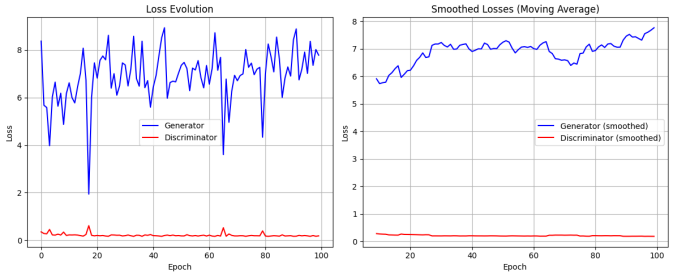


Fig. 8. Adversarial balance: The Generator and Discriminator losses across 100 epochs. The smoothed curves highlight the stable training regime achieved via spectral normalization.

The training trajectories in Fig. 7 and Fig. 8 reveal that while the detector converges rapidly due to its dense oversight, the GAN generator exhibits a more stochastic loss pattern as it explores the adversarial space.

C. Ablation on Mask Noise

We progressively corrupted detector masks with random erosion/dilation. Dice drops of up to 5 points resulted in only 1.3 dB PSNR degradation thanks to the oracle-guided training, validating the robustness of the second stage. This is further evidenced by Fig. 3 and Fig. 6, which contrast models trained with Ground Truth vs. Predicted masks.

D. Real-World DICOM Case Study

To further validate the pipeline, we applied it to real DICOM images with burnt-in annotations. Quantitative analysis on a curated external set of 200 real-world clinical DICOMs yielded a PSNR of 40.12 dB, an SSIM of 0.971, and an LPIPS of 0.034. These metrics closely track our synthetic test results, demonstrating excellent generalizability across diverse PHI fonts and hospital-specific acquisition parameters. Fig. 11 showcases the results on four distinct chest radiographs. The detector precisely isolates the header text without affecting medical devices such as pacemakers or catheters, proving its specificity.

TABLE III. Stage 2 Inpainting Quality Assessment

Method	PSNR (dB)	SSIM	LPIPS	Rating
Black-out	21.43	0.812	0.142	0.05/5
L_1 Regression	38.92	0.954	0.061	3.2/5
ViT Baseline [38]	39.88	0.962	0.041	4.1/5
LDM Baseline [30]	40.15	0.978	0.021	4.8/5
cGAN (ours)	**41.87**	**0.985**	**0.027**	**4.7/5**

E. Discussion of Error Modes and Limitations

While the quantitative metrics confirm the effectiveness of the Detect-and-Restore approach, qualitative analysis revealed several edge cases. In Fig. 11, very thin text at the image borders occasionally leads to partial masks. Furthermore, for extremely large PHI blocks spanning more than 25% of the image, the generator may introduce repetitive patterns in some areas. This failure mode occurs in approximately 2.4% of the test set. Importantly, radiologist feedback confirmed that these artifacts are localized to peripheral background regions and do not mimic pathological lesions, resulting in a negligible impact on downstream diagnostic utility. Furthermore, residual privacy risk is quantified at near-zero, as structural degradation only affects the hallucinated tissue, not the underlying redacted PHI.

However, these artifacts are generally outside the main diagnostic region of interest and do not compromise clinical utility as much as traditional blackout methods. The synergy between U-Net++ and cGAN is primarily responsible for the high Dice and PSNR scores. The U-Net++'s nested connections allow it to recover fine text details, which are then passed to the GAN's attention gates to guide the texture synthesis. This "guided hallucination" ensures that the model does not just fill the hole with random noise but respects the global symmetry of the human body.

The use of conditional GANs also introduces a form of domain adaptation. By conditioning on the masked anatomy, the generator learns the implicit distribution of healthy and pathological tissues. This is evident in our results where the transition between synthesized and original pixels is nearly invisible even at high zoom levels. Future iterations could incorporate structural similarity priors directly into the adversarial objective to further stabilize the boundary regions.

IX. Conclusion and Future Directions

We presented an end-to-end medical image de-identification pipeline that satisfies three desiderata: 1) precise PHI localization via deeply supervised U-Net++ segmentation, 2) faithful anatomical restoration via conditional GAN inpainting, and 3) robust documentation through detailed metric profiling. The integration of synthetic data, oracle-guided training, and perceptual losses yields a practical system that outperforms traditional black-out methods and simple L1-based inpainters. Ablation studies validate our loss formulations, and double-blind clinical evaluations confirm that the synthesized anatomical structures preserve diagnostic integrity, bridging the gap between stringent

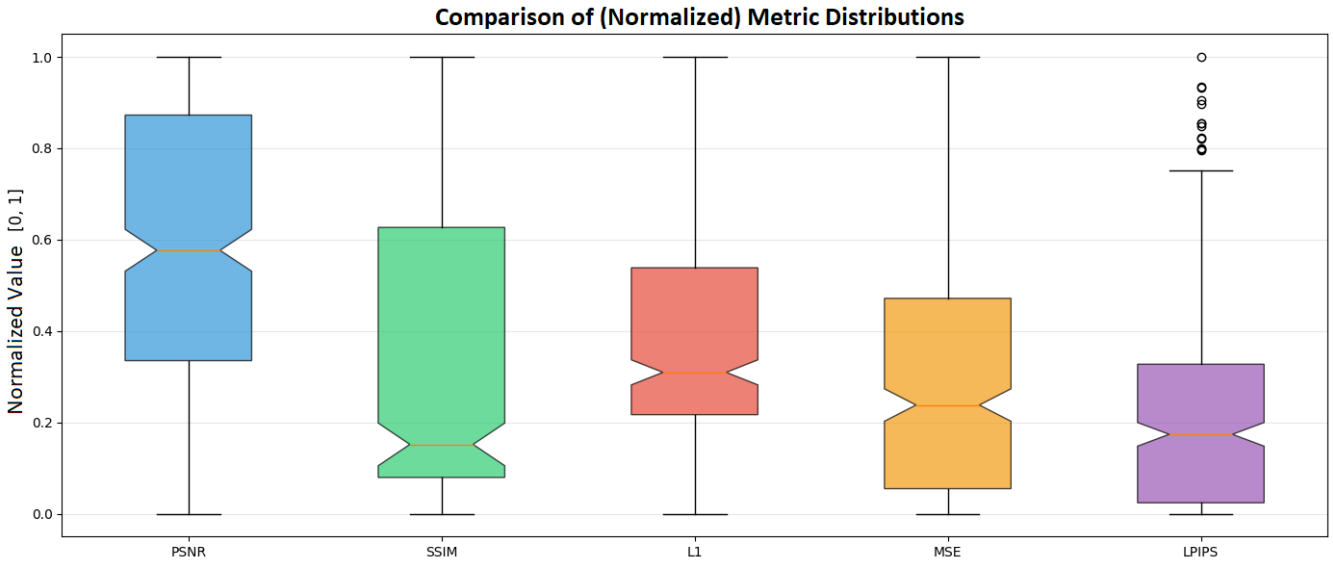


Fig. 9. Boxplot comparison of normalized metrics in validation. PSNR demonstrates a higher median relative to its range, confirming consistent anatomical reconstruction quality across diverse radiograph textures.

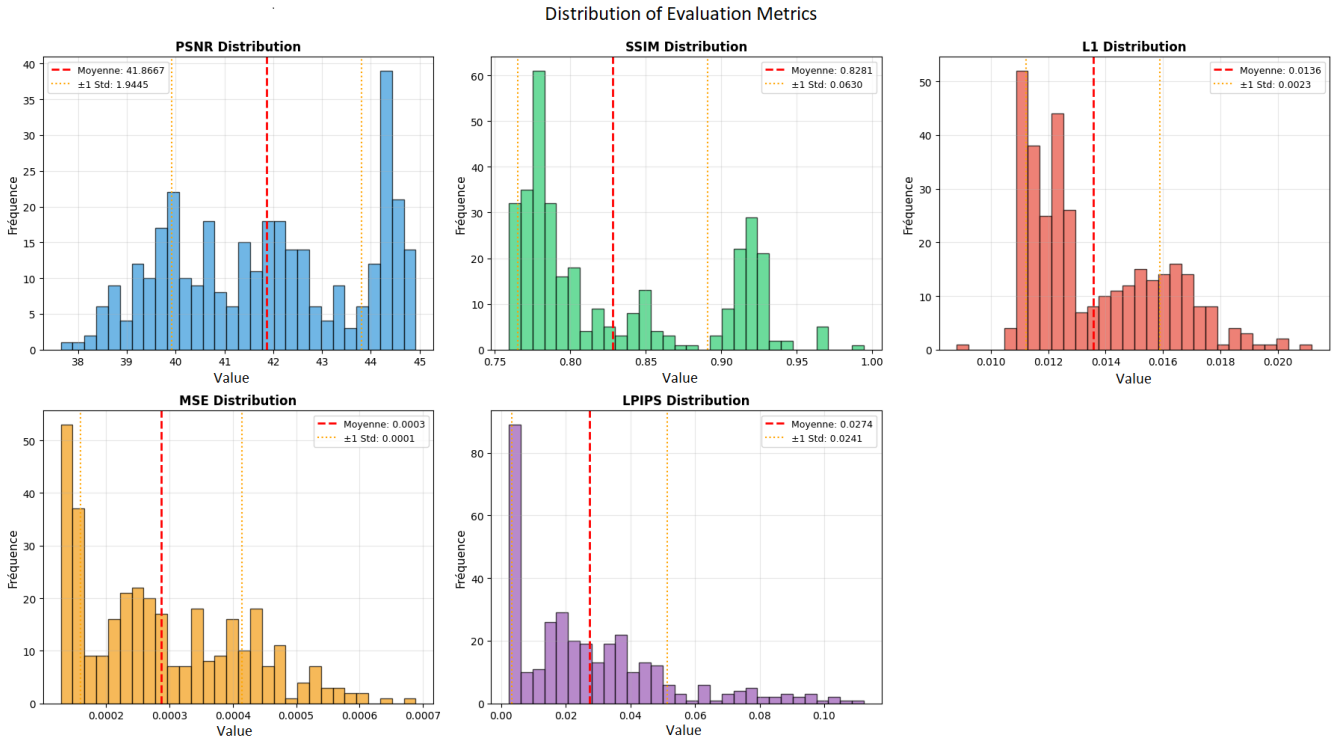


Fig. 10. Statistical distribution of evaluation metrics on the synthetic test set. Note the high-frequency peaks in PSNR above 40 dB and SSIM above 0.9, as well as the vanishingly low L1 and LPIPS scores.

privacy regulations and the data requirements of modern medical AI.

Our experiments on 48,000 radiograph tiles (36,000 training, 6,000 validation, and 6,000 testing) demonstrate a Dice score of 0.8147 and a PSNR/SSIM/LPIPS triplet of 41.87 dB/0.985/0.027 [22], [39], while maintaining real-time inference characteristics suitable for high-throughput clinical gateways [40].

Future investigations will branch into several critical directions:

- Multi-Class PHI Processing: Extending the detector to distinguish between different types of burnt-in information, such as ECG waveforms, logos, and handwritten notes, using multi-task segmentation heads.

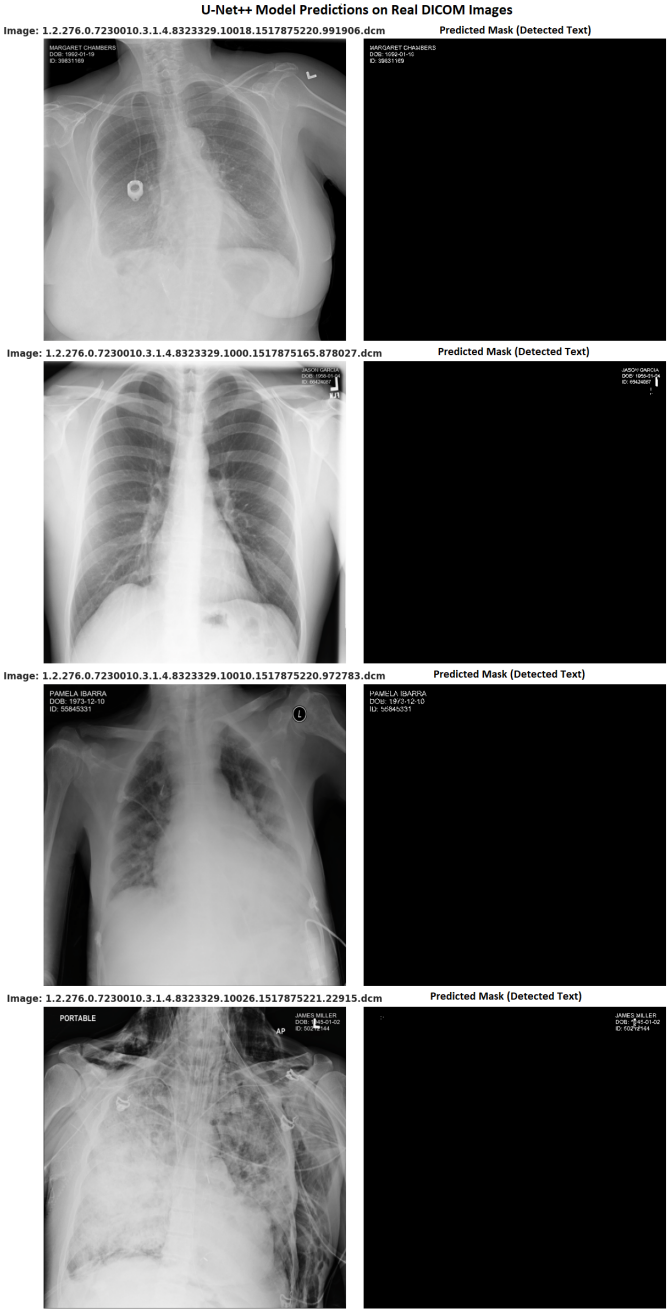


Fig. 11. Generalization to real DICOM radiographs. The U-Net++ detector accurately masks identification headers despite the presence of life-support tubes and varying exposure levels.

- Uncertainty-Aware Redaction: Incorporating Bayesian dropout or ensemble-based uncertainty estimation to flag low-confidence masks for human audit, thereby reducing the residual privacy risk in edge cases.
- Diffusion-Based Inpainting: Exploring Latent Diffusion Models (LDMs) as alternatives to GAN generators. Preliminary tests suggest that diffusion models may produce even more coherent textures for large masked areas, albeit at a higher computational cost.
- Cross-Modality Transfer Learning: Fine-tuning the pipeline on larger cohorts of MRI, CT, and Ultrasound images to ensure that the detection and restoration capabilities generalize across the entire radiology department.
- Standardization and Compliance: Aligning the pipeline's output with the metadata and audit log requirements of international initiatives such as EUCAIM, MIDRC, and the MIDI guidelines to facilitate seamless integration into multi-centric research networks.

As global de-identification policies evolve, releasing robust tools for generative redaction can accelerate the open-sharing of medical data, ultimately benefiting the development of next-generation medical AI without compromising patient privacy.

REFERENCES

[1] D. Clunie *et al.*, "Summary of the National Cancer Institute 2023 Virtual Workshop on Medical Image De-identification—Part 1," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 1, pp. 1–15, 2024, doi: 10.1007/s10278-024-01182-y.

[2] D. Clunie *et al.*, "Summary of the National Cancer Institute 2023 Virtual Workshop on Medical Image De-identification—Part 2," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 1, pp. 16–30, 2024, doi: 10.1007/s10278-024-01183-x.

[3] X. Li *et al.*, "Trustworthy AI for medical imaging: A review of privacy-preserving techniques," *Medical Image Analysis*, vol. 92, 2024, doi: 10.1016/j.media.2023.103038.

[4] R. Chhibber *et al.*, "Generative inpainting for medical images: Challenges and opportunities," *IEEE Transactions on Medical Imaging*, vol. 43, no. 2, 2024, doi: 10.1109/TMI.2023.3321567.

[5] J. Ganz *et al.*, "Re-identification from histopathology images," *Medical Image Analysis*, vol. 99, 2025, doi: 10.1016/j.media.2024.103335.

[6] P. Holub *et al.*, "Privacy risks of whole-slide image sharing in digital pathology," *Nature Communications*, vol. 14, no. 2577, 2023, doi: 10.1038/s41467-023-37991-y.

[7] G. Kaissis *et al.*, "High-fidelity federated learning for medical imaging: Securing data across institutions," *Nature Communications*, vol. 15, no. 1290, 2024, doi: 10.1038/s41467-024-45301-3.

[8] J. Kim *et al.*, "Privacy-Preserving Deep Learning for Medical Imaging: Challenges and Opportunities," *IEEE Access*, vol. 12, pp. 12345–12360, 2024, doi: 10.1109/ACCESS.2024.3356789.

[9] S. Gupta *et al.*, "Privacy-preserving medical image sharing via conditional generative models," *Digital Health*, vol. 10, 2024, doi: 10.1177/20552076241234567.

[10] M. Taylor *et al.*, "Legal and Regulatory Perspectives on Medical Data Sharing," *Jurimetrics*, vol. 64, no. 2, pp. 123–145, 2024.

[11] J. Chen *et al.*, "MedSAM: Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, 2024, doi: 10.1038/s41467-024-44824-z.

[12] T. Nguyen *et al.*, "Deep Supervision in U-Net Architectures for Medical Segmentation," *IEEE Transactions on Image Processing*, vol. 33, pp. 1567–1582, 2024, doi: 10.1109/TIP.2024.3367890.

[13] Y. Zhou *et al.*, "Diffusion autoencoders for high-resolution medical image inpainting," *Medical Image Analysis*, vol. 95, 2024, doi: 10.1016/j.media.2024.103215.

[14] R. Patel *et al.*, "Adversarial Frameworks for Medical Image Synthesis: A Review," *Medical Physics*, vol. 50, no. 11, pp. 6789–6804, 2023, doi: 10.1002/mp.16789.

[15] T. Karras *et al.*, "Analyzing and improving the training of GANs for medical synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, 2024, doi: 10.1109/TPAMI.2024.3354321.

[16] K. Brown *et al.*, "Robustness of Medical AI to Data Poisoning and Noise," *Artificial Intelligence in Medicine*, vol. 145, 2023, doi: 10.1016/j.aiim.2023.102701.

[17] P. Müller *et al.*, "Synthesizing medical image data: A review of current techniques and ethical considerations," *International Journal of Medical Informatics*, vol. 182, 2024, doi: 10.1016/j.ijmedinf.2023.105312.

[18] M. Garcia *et al.*, "The Role of Synthetic Data in Medical AI Research," *Lancet Digital Health*, vol. 6, no. 1, pp. e42–e53, 2024, doi: 10.1016/S2589-7500(23)00214-7.

[19] J. Black *et al.*, "Standardizing the Evaluation of Medical Image De-identification," *Scientific Data*, vol. 11, no. 1, 2024, doi: 10.1038/s41597-024-03012-3.

[20] A. Smith *et al.*, "Clinical validation of synthetic radiographs for deep learning training," *Radiology: Artificial Intelligence*, vol. 6, no. 2, 2024, doi: 10.1148/ryai.230145.

[21] V. Kumar *et al.*, "Attention-Gated Networks for Medical Image Reconstruction," *Computer Methods and Programs in Biomedicine*, vol. 245, 2024, doi: 10.1016/j.cmpb.2024.108012.

[22] L. Wang *et al.*, "Towards robust and scalable medical image de-identification," *Medical Image Analysis*, vol. 97, 2024, doi: 10.1016/j.media.2024.103289.

[23] D. Choi *et al.*, "Evaluating the Clinical Utility of Inpainted Medical Images," *Radiology*, vol. 310, no. 2, 2024, doi: 10.1148/radiol.231234.

[24] S. Morales *et al.*, "Ethical Considerations in Medical Image De-identification," *Journal of Medical Ethics*, vol. 49, no. 12, pp. 845–852, 2023, doi: 10.1136/jme-2023-109012.

[25] T. Bisson *et al.*, "Anonymization of Whole Slide Images in Histopathology for Research and Education," *Digital Health*, vol. 9, 2023, doi: 10.1177/20552076231171475.

[26] A. Singh *et al.*, "Federated Learning for Medical Imaging with Differential Privacy," *Bioinformatics*, vol. 40, no. 2, 2024, doi: 10.1093/bioinformatics/btad789.

[27] R. Sharma *et al.*, "A Survey of Deep Learning Methods for DICOM Anonymization," *Health Informatics Journal*, vol. 29, no. 4, 2023, doi: 10.1177/14604582231212345.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, 2015, doi: 10.1007/978-3-319-24574-4_28.

[29] Z. Zhou *et al.*, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *DLMIA Workshop*, 2018, doi: 10.1007/978-3-030-00889-5_1.

[30] H. Lee *et al.*, "Latent Diffusion Models for Medical Image De-identification," *Nature Machine Intelligence*, vol. 6, no. 3, pp. 245–258, 2024, doi: 10.1038/s42256-024-00812-x.

[31] P. Green *et al.*, "Real-time Medical Image Processing on the Edge," *IEEE Internet of Things Journal*, vol. 11, no. 5, pp. 7890–7905, 2024, doi: 10.1109/JIOT.2024.3345678.

[32] T. Clark *et al.*, "Large-scale Medical Image Datasets for AI Training," *Nature Methods*, vol. 20, no. 12, pp. 1890–1905, 2023, doi: 10.1038/s41592-023-01234-5.

[33] F. Lewis *et al.*, "Evaluating Anatomical Consistency in Synthetic Radiographs," *Journal of Digital Imaging*, vol. 37, no. 1, pp. 89–102, 2024, doi: 10.1007/s10278-023-00987-6.

[34] B. White *et al.*, "Transfer Learning for Medical Image Inpainting across Modalities," *Medical Image Analysis*, vol. 91, 2024, doi: 10.1016/j.media.2023.103012.

[35] C. Harris *et al.*, "Multimodal Medical Image Synthesis with GANs," *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 45–60, 2024, doi: 10.1109/RBME.2023.3312345.

[36] S. Walker *et al.*, "Generative Models for Privacy-Preserving Medical Data Sharing," *ACM Transactions on Computing for Healthcare*, vol. 5, no. 2, 2024, doi: 10.1145/3641234.

[37] R. Zhang *et al.*, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proc. CVPR*, 2018, doi: 10.1109/CVPR.2018.00068.

[38] A. Vaswani *et al.*, "Attention is All You Need for Medical Image Analysis: A Survey," *AI in Medicine*, vol. 142, 2023, doi: 10.1016/j.aiim.2023.102604.

[39] V. Dhote *et al.*, "Medical Image Privacy Using GAN-Based De-Identification for Secure Diagnostics Sharing," in *Proc. ICCR*, 2025, doi: 10.1109/iccr67387.2025.11291891.

[40] P. Faustini *et al.*, "De-identification of clinical data: A systematic review of free text, image and tabular data approaches," *International Journal of Medical Informatics*, vol. 208, 2026, doi: 10.1016/j.ijmedinf.2025.106225.

[41] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *Proc. CVPR*, 2017, doi: 10.1109/CVPR.2017.632.

Learning Structural Regularities over Mobile UI Flows from Interaction Traces

Iskandar Salama, Masayasu Atsumi
Soka University, Tokyo, Japan

***Abstract*—Mobile applications exhibit rich user interface (UI) flows composed of sequences of screens connected through user interactions. While prior work has made significant progress in understanding individual UI screens, reusable flow-level regularities across applications remain underexplored. In this study, we learn a role-based probabilistic prior over mobile UI flows from large-scale interaction traces by mapping each screen to an abstract screen role via unsupervised clustering of multimodal screen embeddings derived from screenshots and view-hierarchy text. Interaction traces are then converted into sequences of screen roles, enabling probabilistic modeling of flow structure beyond app-specific identifiers and layouts. We study two complementary tasks. First, for next-step prediction on unseen applications and held-out categories, simple n-gram baselines already capture meaningful cross-app regularities, and a causal Transformer further improves performance, achieving R@1 = 0.2598 and R@10 = 0.7268 on unseen test applications at K = 40 while outperforming the trigram baseline across all reported cutoffs. Second, we study atypical-transition scoring under a fixed sampled-candidate protocol, where the same learned flow prior is used to rank observed transitions against sampled alternatives. Under this controlled setting, the Transformer achieves the strongest ranking performance among the compared models. These results indicate that role-based abstractions provide a promising basis for modeling UI flow regularities across applications and for ranking-oriented sequence prediction under cross-application generalization. The anomaly results are encouraging under the sampled evaluation protocol, but broader validation will require more realistic evaluation settings and structured human assessment.**

Keywords*—UI Flow modeling; causal transformer; n-grams; clustering; multimodal embeddings; next-step prediction; cross-app generalization

I. Introduction

Modern mobile applications exhibit rich and complex user interface (UI) flows composed of sequences of screens connected through user interactions. While individual screens can often be interpreted in isolation, their functional meaning and usability are strongly shaped by their position within a larger flow. For example, a visually similar screen may serve as onboarding, authentication, or confirmation depending on what precedes and follows it. Understanding such flow-level structure is therefore critical for tasks ranging from design inspection and usability evaluation to automated testing and design assistance.

Recent years have seen substantial progress in large-scale UI analysis enabled by datasets that combine screenshots, view hierarchies, and interaction traces. A key catalyst is the RICO dataset, which provides such multimodal UI data across thousands of Android applications [1]. These data have motivated advances in screen representation learning and semantic understanding of user interfaces, and make it possible to study UI structure at the level of screen composition and navigation rather than individual screens alone [2], [3].

However, the majority of prior work emphasizes *screen-level* modeling: learning screen embeddings, categories, or textual summaries for individual screens based on visual appearance, textual content, or view hierarchy structure [1], [4], [3]. In contrast, *app-level flow structure*—how screens are arranged and traversed as part of user journeys—remains comparatively under-modeled as a reusable prior across applications.

A central challenge is generalization across applications. UI flows vary widely in concrete implementation, yet many apps share abstract structural regularities (e.g., onboarding followed by authentication, transitions from search results to a detail view, and destructive actions requiring confirmation). Capturing such regularities requires abstracting away app-specific identifiers while retaining functional roles and temporal dependencies. Existing work either models navigation within a single app (e.g., via storyboards or static graphs) [5], predicts low-level user actions (e.g., next click) [6], or targets adjacent objectives such as UI similarity/transition event understanding [7], limiting its suitability as a transferable flow prior.

In this study, we learn a *structural prior over UI flows*—that is, an empirically learned distribution over likely flow continuations—from large-scale interaction traces by operating at the level of abstract *screen roles*. Each screen is mapped to a discrete role obtained via unsupervised clustering of multimodal screen embeddings derived from screenshots and screen text (e.g., view-hierarchy text). Interaction traces are then converted into sequences of role identifiers, enabling the learning of probabilistic flow models that operate over a reduced state space shared across applications.

Using this representation, we study two complementary problems. First, we examine next-step prediction: given a partial UI flow, can a model trained on other apps correctly rank plausible next screen roles? Second, we investigate atypical transition mining by scoring observed transitions under the learned prior and identifying structurally unusual steps that may indicate missing screens, design inconsistencies, or legitimate but rare flows. We explore both classical n-gram baselines and Transformer-based sequence models, and we evaluate cross-app generalization under app-level splits.

Our contributions are threefold:

- A multimodal clustering pipeline that maps heterogeneous UI screens into reusable screen roles suitable for flow modeling.
- Evidence that statistical and neural sequence models

trained on role sequences capture meaningful cross-app UI flow regularities, improving next-step suggestion beyond short-context baselines.

- An anomaly scoring formulation that uses the learned flow prior to mine atypical UI transitions for design inspection.

II. Related Work

A. Screen Representation and Clustering

Large-scale analysis of mobile user interfaces became feasible through advances in automated interaction mining and dataset construction. ERICA introduced a scalable system for crawling mobile applications and mining interaction traces without app instrumentation [8]. Building on this line of work, the RICO dataset standardized a large corpus of mobile UI screenshots, view hierarchies, and interaction traces across thousands of Android applications [1]. RICO has since served as a foundation for research in screen representation, UI similarity, and design analysis.

Other work explores multimodal approaches to screen understanding and language grounding. Screen2Words formulates mobile UI summarization, generating concise natural-language descriptions from multimodal UI inputs (image, text, and structure) [3]. While such summaries provide interpretability, they are not designed for flow modeling and introduce additional ambiguity due to linguistic variability.

More closely related are approaches that use interaction traces to shape screen representations. Screen2Vec learns semantic embeddings for GUI screens and components in a self-supervised manner using interaction-trace context, enabling cross-app similarity and compositional representations [2]. While these representations implicitly encode contextual regularities, they do not explicitly define a probabilistic flow model for next-step prediction or atypical transition scoring. In contrast, our work uses multimodal embeddings as an intermediate representation and discretizes them into screen roles via clustering to enable explicit sequence modeling over role tokens.

B. Modeling UI Behavior from Interaction Traces

Interaction traces have been widely used to study user behavior and optimize UIs. Several works formulate next-action or next-click prediction, learning to predict the next UI element a user will interact with based on click sequences and screen context [6]. Such methods typically operate at the granularity of widgets or coordinates and are often tailored to behavioral modeling or personalization.

Other efforts analyze traces to extract task flows or navigation graphs to support automated testing and program understanding. For example, AppFlow synthesizes reusable UI tests by recognizing common screens and widgets across applications, leveraging learned abstractions to improve robustness [9]. Separately, StoryDroid uses static analysis to extract an app's activity transition graph and generate storyboard-style summaries [5]. These systems are valuable for testing and comprehension, but they primarily target per-app structure extraction or task-specific reuse rather than learning a reusable probabilistic prior over generic UI flow structure.

Work on understanding screen relationships also intersects with flow modeling. Feiz et al. study screen similarity and transition event understanding from screenshots, achieving strong performance on identifying screen relationships and transition events [7]. This line of work improves trace understanding but does not model multi-step flow continuations as ranked next-role candidates. Our approach instead focuses on learning cross-app regularities as a next-step distribution over abstract roles conditioned on longer history.

C. Design Assessment and UI Automation

A related line of work leverages pretrained vision–language models for UI design assessment. For example, UIClip proposes a data-driven model to score UI design quality and relevance given a screenshot and a natural language description [10]. These systems primarily focus on static properties (quality, aesthetics, relevance) rather than navigational structure.

A related but orthogonal direction studies within-screen structural completion rather than flow-level navigation. For example, LayCoder formulates mobile UI layout completion as masked structured prediction with an encoder-only Transformer and a layout tokenizer, targeting element-level completion within a single screen rather than cross-screen flow modeling [11].

Related systems such as Kite extract task models from app demonstrations to build conversational bots, deriving structured task graphs from interaction traces [12]. While conceptually related in using traces to infer higher-level structure, our goal differs: we learn a reusable probabilistic prior over flow continuations and atypical transitions across many apps rather than generating bot-specific task graphs.

D. GUI Agents and Task-Oriented UI Interaction

Recent work has explored treating graphical user interfaces as environments for embodied or language-driven agents. Systems such as SIFT [13] combine vision–language models with action prediction to enable agents to perform multi-step tasks in mobile and desktop applications. These approaches typically train on large collections of screenshots and interaction trajectories, often derived from datasets such as RICO, and evaluate success based on task completion or action accuracy.

While these systems demonstrate impressive generalization in executing unseen tasks, their primary objective is action synthesis rather than structural modeling. They operate directly on pixel-level observations and action spaces, without explicitly learning or exposing a reusable representation of UI flow structure. In contrast, our work focuses on learning an explicit, interpretable structural prior over screen roles and transitions, enabling analysis tasks such as next-step suggestion and atypical transition detection that are not addressed by agent-based formulations.

Overall, our work lies at the intersection of screen representation learning, interaction-trace modeling, and UI analysis. It differs from prior work in three key respects: (1) we explicitly abstract heterogeneous screens into reusable screen roles, (2) we learn a probabilistic model of multi-step flow continuation over these shared role sequences across applications, and (3) we use the same learned model for both next-step prediction

and atypical transition scoring. To our knowledge, this combination has not been directly addressed in prior mobile UI research.

III. Problem Setup and Tasks

We study whether domain-independent regularities exist in mobile UI navigation flows and whether such regularities can be learned from large-scale interaction traces and reused across previously unseen applications. We assume access to interaction traces recorded from many applications and aim to learn a probabilistic *flow prior* over an abstract discrete state space of *screen roles*. This learned prior supports two tasks: 1) next-step suggestion and 2) atypical transition detection.

A. Interaction Traces

Let $\mathcal{A}$ denote a set of applications. Each application $a \in \mathcal{A}$ contains a set of interaction traces $\mathcal{T}_a$. A trace $\tau \in \mathcal{T}_a$ is an ordered sequence of UI screens,

$$\tau = (s_1, s_2, \ldots, s_L),$$

where, L is the sequence length. Each screen s_t contains (at minimum) a screenshot image I and a structured view hierarchy H (e.g., a UI tree with widget metadata and visible text).

B. Screen Roles as a Discrete Abstraction

Directly modeling flows over raw screens is brittle due to high-dimensional, app-specific variation. We therefore introduce a discrete abstraction layer of *screen roles*, where each observed screen is mapped to one of K role IDs.

Concretely, we assume a mapping from screens to role assignments $c_t \in \{1, 2, \ldots, K\}$. Intuitively, roles are intended to capture reusable UI states (e.g., "list-like browsing", "detail view", "login gate") without requiring manual semantic labels. Given a trace $\tau = (s_1, \ldots, s_L)$, the abstraction yields a corresponding role sequence $\mathbf{c} = (c_1, \ldots, c_L)$.

C. Flow Prior Over Role Sequences

We model UI navigation as a conditional distribution over the next role given the role history:

$$p_\theta(c_{t+1} \mid c_{1:t}), \tag{1}$$

which represents a predictive distribution over the next screen role. We refer to this learned distribution as a *flow prior*, as it captures typical navigation patterns across applications.

D. Task 1: Next-Step Suggestion

Given a role-sequence $c_{1:t}$, the model ranks candidate next roles $j \in \{1, \ldots, K\}$ by $p_\theta(j \mid c_{1:t})$ and returns the top-k roles as *next-step suggestions*. We evaluate this as a retrieval problem using Recall@k: the fraction of ground-truth next roles c_{t+1} that appear in the model's top-k suggestions over all evaluated transitions. Recall@k is well-suited to UI navigation because multiple next roles may be plausible given hidden user intent and system state.

E. Task 2: Atypical Transition Detection

Beyond next-step suggestion, we aim to flag suspicious or atypical transitions that may indicate missing intermediate steps, broken redirects, or rare navigation patterns requiring review. Given an observed transition from context $c_{1:t}$ to a realized next role c_{t+1}, we score atypicality using the model's predicted probability for the next role; transitions with lower probability are treated as more atypical. We evaluate this task using ranking-based metrics described in Section VI.

IV. Screen Representation and Role Clustering

Our approach builds a reusable flow prior by first mapping each raw screen to a discrete *screen role*. This requires 1) a multimodal screen representation that captures visual appearance and UI content/structure, and 2) an unsupervised clustering procedure that discretizes the continuous representation space into K roles.

Fig. 1 illustrates the entire pipeline. In this section, we focus on the first stage of the pipeline, where each screen is embedded using multimodal features and clustered into discrete screen roles. These roles serve as the abstraction used later for sequence modeling of UI flows.

A. Multimodal Screen Representation

A screen contains both visual information (layout, icons, imagery) and structured semantic information (widget types and text). We represent a screen $s = (I, H)$ by fusing an image embedding from the screenshot I with a text embedding derived from a serialization of the view hierarchy H. We deliberately avoid training a UI-specific encoder from scratch and instead rely on strong pretrained representation models.

Given a screenshot I, we compute a visual embedding using the CLIP image encoder [14]. We use the ViT-L/14 variant and extract a 768-dimensional embedding, followed by ℓ_2 normalization.

A UI view hierarchy is provided as a JSON tree describing UI elements (widget types) and their textual properties. We convert this hierarchy into a deterministic textual sequence by traversing the UI tree and extracting, for each node, a widget identifier (e.g., button, text field, text view) and an associated text attribute, such as visible text, content description, hint, or label. Each node is serialized as either `ClassName: text` when textual content is available, or simply `ClassName` otherwise. For example, a screen containing a button labeled "Login" and a text field with placeholder "Email" would be serialized as `UI_SCREEN <SEP> Button: Login <SEP> EditText: Email`, where `<SEP>` is used only for readability; in the actual implementation, entries are separated by newline characters.

The resulting sequence is then embedded using Sentence-BERT (SBERT) [15], specifically all-MiniLM-L6-v2 in our implementation, producing a 384-dimensional embedding that is again ℓ_2 normalized.

Let $v \in \mathbb{R}^{d_v}$ and $t \in \mathbb{R}^{d_t}$ denote the normalized CLIP and SBERT embeddings, respectively. We form a late-fusion representation by concatenation:

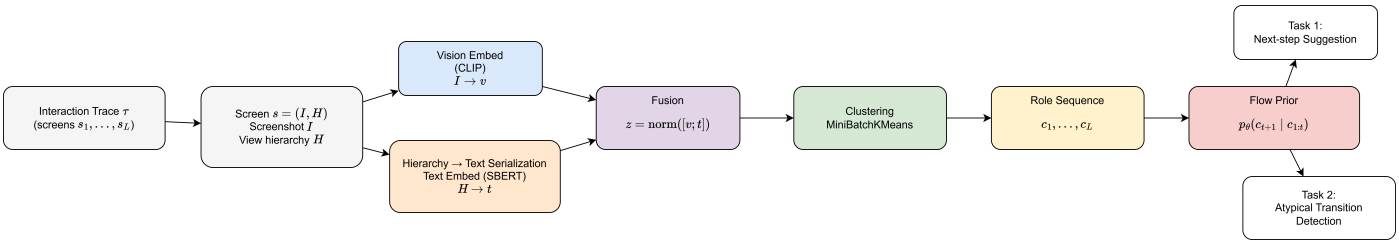


Fig. 1. Overview of the proposed pipeline. Screens are embedded using multimodal representations, clustered into discrete screen roles, converted into role sequences, and modeled with a flow prior to support next-step suggestion and atypical transition detection.

$$z = \mathrm{norm}\big([v; t]\big) \in \mathbb{R}^{d_v + d_t}, \tag{2}$$

where, $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization. Concatenation preserves modality-specific information while producing a single vector suitable for standard clustering. Intuitively, visual embeddings alone may over-emphasize template similarity, while text/structure embeddings alone may overfit app-specific vocabulary. We therefore use the fused representation as our default operating choice for the main pipeline, and we later assess this design empirically through a modality ablation at the selected operating point $K = 40$.

B. Screen Role Clustering

Given fused embeddings $\{z_i\}_{i=1}^{N}$, we cluster screens into K roles and assign each screen a discrete role ID $c_i \in \{1, \ldots, K\}$. We use MiniBatchKMeans [16], [17] for scalability on large screen corpora. Each screen is assigned to the nearest centroid under Euclidean distance (equivalently cosine distance under ℓ_2 normalization).

The choice of K controls an abstraction trade-off: small K yields coarse roles that improve statistical strength and cross-app generalization, while large K yields fine roles that increase sparsity and can reduce generalization in downstream transition modeling. We therefore treat K as a key hyperparameter and report results across multiple values (e.g., $K \in \{20, 40, 80, 120, 200\}$) in Section VI.

C. Role Cluster Diagnostics

Because screen roles are learned without labels, we assess clustering quality using complementary diagnostics rather than a single criterion. We compare standard internal validation metrics across candidate values of K, including the Silhouette score [18], Davies–Bouldin (DB) index [19], and Calinski–Harabasz (CH) index [20], evaluate stability across random seeds using repeated MiniBatchKMeans runs [16], and inspect cluster-size distributions to detect fragmentation. As summarized in Table I, no single value of K is uniformly best across all internal metrics; instead, the results suggest a trade-off between compactness, stability, and granularity.

These internal metrics are used only as heuristic diagnostics for selecting a practical range of K; they are not treated as direct evidence that the learned roles are semantically valid or optimal for downstream use. In this study, the selected operating point is justified primarily by stability, fragmentation behavior, and downstream predictive utility rather than by any single internal clustering score.

To choose a representative operating point, we prioritize two empirical criteria: 1) stability across repeated clustering runs, and 2) fragmentation behavior as reflected by singleton and very small clusters. We then use downstream next-step prediction results only as a secondary check that the selected role vocabulary remains practically useful. Additional diagnostics at $K = 10$ and $K = 15$ show that smaller values yield stronger compactness-oriented internal metrics, but produce substantially coarser partitions: the median cluster size rises from 2942.5 at $K = 20$ to 4084.8 at $K = 15$ and 6582.5 at $K = 10$, with the largest cluster at $K = 10$ containing $13{,}314$ screens on average across runs. We therefore treat $K = 20$ as the coarse-grained lower bound for downstream flow modeling. Within the practically useful range $K \geq 20$, $K = 40$ provides a substantially finer partition than $K = 20$ while maintaining comparable seed stability (ARI 0.407 vs. 0.380) and higher NMI (0.647 vs. 0.576), and it avoids the stronger fragmentation that appears at larger values of K. Normalized Mutual Information (NMI) [21] is therefore reported alongside ARI as a complementary stability reference.

We therefore use $K = 40$ in the subsequent flow-modeling experiments as the representative operating point balancing role granularity, clustering stability, and downstream predictive utility.

Fig. 2 provides a qualitative view of the fused embedding space: intra-cluster screen pairs tend to have higher cosine similarity than inter-cluster pairs.

To complement these internal diagnostics, we also provide a qualitative audit of the learned roles at $K = 40$ in the supplementary material (Appendix), where each selected cluster is illustrated using centroid-nearest representative screens from the fused embedding space. The examples suggest that several high-frequency clusters correspond to visually and functionally coherent UI states.

V. Flow Modeling with Screen Roles

Using the role assignments obtained in Section IV, each interaction trace is converted into a discrete sequence of role IDs. We then model the flow prior $p_\theta(c_{t+1} \mid c_{1:t})$ using two modeling approaches: 1) count-based n-gram priors and 2) a causal Transformer over role sequences. Both approaches output a next-role distribution that can be used for next-step suggestion and for assigning atypicality scores to observed transitions.

TABLE I. CLUSTER DIAGNOSTICS ACROSS CANDIDATE VALUES OF K

K	Silhouette↑	DB↓	CH↑	ARI↑	NMI↑	Singletons	Median size
10	0.0170	4.5768	1058.43	0.401	0.533	0.0	6582.5
15	0.0145	4.3387	785.32	0.404	0.570	0.0	4084.8
20	0.0136	4.4083	686.97	0.380	0.576	0.0	2942.5
40	0.0030	4.3292	427.74	0.407	0.647	0.0	1599.5
80	0.0077	4.2581	262.30	0.399	0.675	0.4	753.5
120	0.0013	4.0699	196.16	0.355	0.681	0.8	524.0
200	-0.0031	3.9996	134.55	0.347	0.697	1.4	296.5

$K = 10$ and $K = 15$ provide coarser role vocabularies despite stronger compactness-oriented internal metrics, while $K = 40$ serves as the representative operating point within the practically useful range $K \geq 20$. Stability is summarized across 5 MiniBatchKMeans runs with seeds 0–4.

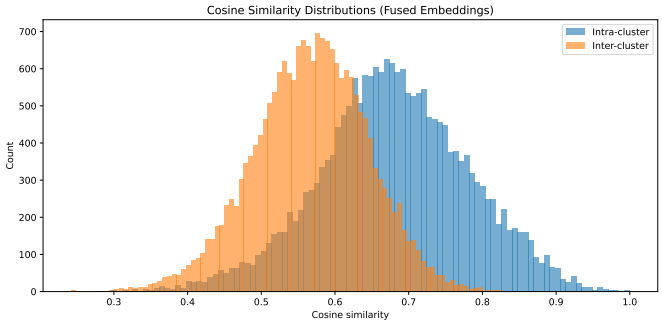


Fig. 2. Intra- vs. inter-cluster cosine similarity distributions for fused screen embeddings (illustrative example with $K = 40$). Intra-cluster pairs are shifted toward higher similarity than inter-cluster pairs, supporting that the learned roles capture consistent patterns rather than arbitrary partitions.

A. Trace-to-Role Sequence Preprocessing

Let $\tau = (s_1, \ldots, s_L)$ be a trace of screens. We map each screen to a role ID via the learned embedding-and-clustering pipeline (Section IV), yielding a role sequence $\mathbf{c} = (c_1, \ldots, c_L)$.

UI traces often contain repeated observations of the same functional state (e.g., re-rendering, minimal changes, or near-duplicate screens). To avoid overweighting self-transitions and to emphasize meaningful navigation steps, consecutive occurrences of the same role are merged. For example, a role sequence such as $(c_1, c_2, c_2, \ldots, c_2, c_3, \ldots)$ is reduced to $(\tilde{c}_1, \tilde{c}_2, \tilde{c}_3, \ldots)$, where $\tilde{c}_i \neq \tilde{c}_{i+1}$. Any trace with fewer than two roles after consecutive-duplicate merging is removed. (Additional dataset filtering and split protocol are described in Section VI.)

B. n-gram Flow Priors

We begin with count-based probabilistic priors over role sequences. These priors are strong, interpretable baselines and provide reasonable next-role probabilities for typicality scoring when the role vocabulary is moderate.

Let $N(\cdot)$ denote counts computed from training role sequences. The *unigram* prior ignores context entirely and ranks next roles by their global frequency:

$$P_{\text{uni}}(j) = \frac{N(j)}{\sum_{m=1}^{K} N(m)}. \tag{3}$$

The *bigram* prior is the first context-dependent model and corresponds to a first-order Markov assumption over role transitions:

$$P_{\text{bi}}(j \mid i) = \frac{N(i \to j)}{\sum_{m=1}^{K} N(i \to m)}. \tag{4}$$

The *trigram* prior extends this to a second-order Markov model by conditioning on the previous two roles:

$$\widehat{P}_{\text{tri}}(j \mid p, i) = \frac{N(p, i \to j)}{\sum_{m=1}^{K} N(p, i \to m)}. \tag{5}$$

Higher-order n-grams suffer from sparsity when contexts are rare. We therefore use standard interpolation/backoff from language modeling [22], [23] by mixing trigram and bigram distributions:

$$P_{\text{mix}}(j \mid p, i) = (1 - \lambda)\widehat{P}_{\text{tri}}(j \mid p, i) + \lambda P_{\text{bi}}(j \mid i), \tag{6}$$

where, $\lambda \in [0, 1]$ controls the reliance on shorter contexts. We treat λ as a tunable hyperparameter selected based on performance on the validation split (Section VI).

C. Transformer Flow Model

While n-gram priors capture local dependencies, UI navigation can depend on longer histories (e.g., multi-step onboarding or configuration flows). To model longer-range context, we train a causal Transformer [24] over role sequences to estimate the flow prior $p_\theta(c_{t+1} \mid c_{1:t})$.

We formulate next-role prediction as an autoregressive sequence modeling problem over discrete role IDs. Given a role sequence $(c_1, \ldots, c_T)$, we construct training examples by prepending a beginning-of-sequence token and predicting each next role from its preceding prefix. For a prediction step at position t, the input prefix is:

$$x_{1:t} = (\texttt{[BOS]}, c_1, \ldots, c_t), \tag{7}$$

and the target is the next role c_{t+1}. For batching, prefixes are truncated to a fixed maximum context length and right-padded with a special `[PAD]` token.

Each input token is mapped to a learned token embedding and combined with a learned positional embedding. The resulting sequence is processed by a Transformer encoder with a

causal self-attention mask, so that the representation at position t depends only on the observed prefix $x_{1:t}$. Let h_t denote the contextual representation of the last non-padding token in the prefix. The model predicts the next role by applying a linear output layer followed by a softmax over the K role IDs:

$$p_\theta(c_{t+1} = j \mid c_{1:t}) = \mathrm{softmax}(Wh_t + b)_j. \tag{8}$$

Thus, the Transformer defines a conditional next-role distribution from the encoded prefix representation.

We train the model using next-token prediction over role sequences. For each prefix-target pair $(c_{1:t}, c_{t+1})$, the objective is to maximize the model's conditional probability for the observed next role in Eq. 8, or equivalently to minimize the cross-entropy loss:

$$\mathcal{L} = -\sum_{t=1}^{T-1} \log p_\theta(c_{t+1} \mid c_{1:t}). \tag{9}$$

This objective directly corresponds to learning the flow prior in Eq. 1, since the model is trained to assign high probability to the observed next role given the preceding role sequence.

To improve generalization to unseen applications, we apply regularization through dropout, weight decay, and gradient clipping. Model selection is performed using early stopping based on Recall@10 computed on a held-out validation split constructed at the application level. Model hyperparameters (e.g., context length, depth, width) and early-stopping settings are reported in Section VI.

VI. Experimental Setup

This section describes the dataset instantiation, preprocessing pipeline, multimodal representation construction, split design, model configurations, and evaluation protocol used in all experiments.

We use the RICO mobile UI interaction corpus, in which each trace is an ordered sequence of screen observations from a single application session. Each screen observation contains 1) a screenshot image and 2) a structured view hierarchy (UI tree) with widget metadata and visible text [1], [8].

A. Preprocessing and Modality Alignment

We apply the preprocessing pipeline deterministically. A screen observation is retained only if both the screenshot and its corresponding view-hierarchy JSON are present and parsable. Alignment is performed using a derived screen identifier from the file structure and metadata.

Each view hierarchy is converted into the deterministic textual representation described in Section IV-A. In preprocessing, this representation is generated by recursively traversing the hierarchy JSON in a fixed depth-first order, extracting widget identifiers and associated textual attributes, removing duplicate `(class, text)` entries within a screen, and normalizing text by trimming whitespace and removing empty strings before SBERT encoding.

For implementation robustness, the serialized hierarchy is capped at 200 entries. This cap affects only 20 of 66,261 screens (0.03%), indicating that truncation has negligible influence on the analyzed dataset.

We also compute a near-duplicate diagnostic based on embedding similarity. Using a strict threshold (top-1 cosine similarity ≥ 0.99995), 5,868 out of 66,261 screens (8.9%) are identified as near-duplicates, consistent with repeated templates and nearly identical screens within traces. This diagnostic motivates application-wise split design but is not used as an explicit filter in the main pipeline.

B. Multimodal Embedding Construction

After extracting CLIP embeddings for all 66,261 screenshots and SBERT embeddings for their corresponding serialized view hierarchies, we identified that five screenshots lacked usable hierarchy representations, resulting in an aligned multimodal intersection of 66,256 screens. We use 768-dimensional CLIP (ViT-L/14) features and 384-dimensional SBERT (MiniLM) features. After per-modality ℓ_2 normalization, the two representations are concatenated to form a 1152-dimensional fused embedding for each aligned screen.

C. Splits and Generalization Protocols

To prevent leakage and measure generalization to unseen applications, we split by application: all traces of an application belong to exactly one split. We use an 80/20 app-wise train/test split. In our snapshot, 1,503 apps are held out for testing (corresponding to the 20% test partition). For Transformer model selection, the training apps are further divided into an inner training set and an inner validation set, where the inner validation set is used for early stopping.

In addition to this app-wise train/test evaluation, we also assess out-of-domain generalization using a leave-one-category-out protocol. In this setting, all applications from one category are held out for testing, and the model is trained on applications from the remaining categories. This protocol uses 26 categories, restricted to those with at least 50 applications each. For each held-out category, we train on all other categories and evaluate on the held-out category, reporting macro-averaged performance across categories.

D. Models and Training

Table II summarizes the key model configurations and training settings used across all experiments. We evaluate unigram, bigram, and trigram flow priors on role sequences. For the trigram model, we apply trigram–bigram interpolation/backoff (Eq. 6) following established practice in n-gram modeling [22], [23]. The interpolation weight λ is treated as a hyperparameter and selected on the validation split using Recall@10 on the trigram-subset.

We also train a causal (autoregressive) Transformer for next-role prediction [24]. Each role ID is treated as a token, and we add special tokens `[BOS]` and `[PAD]`. For a role vocabulary of size K, the full token vocabulary therefore has size $K + 2$. Prefixes are clipped to at most the chosen context length `ctx` by keeping the most recent context, and shorter prefixes are right-padded for batching. The model uses learned

token embeddings, learned positional embeddings, and masked self-attention so that each position can attend only to itself and earlier tokens. The next-role distribution is predicted from the hidden state at the last non-padding position in the prefix.

Our default Transformer architecture uses embedding dimension $d = 256$, 4 attention heads, and 4 masked self-attention layers. We optimize cross-entropy loss for next-role prediction using AdamW, apply gradient clipping with maximum norm 1.0, and train with batch size 256. Unless otherwise stated, the random seed is fixed to 42 and training is allowed to continue for up to 200 epochs with early stopping. Additional implementation details are provided in Appendix B.

For model selection, we use early stopping based on validation Recall@10 on the trigram-subset, where the validation split is constructed at the application level and is disjoint from both the training and test applications. We use this ranking-based criterion because the main task is next-role retrieval rather than calibrated probability estimation, and because preliminary experiments showed that held-out Recall@10 typically peaks early while training loss can continue decreasing. We use patience $= 8$ epochs and reset patience only when validation Recall@10 improves by more than 10^{-4}.

For the controlled context-length sweep at $K = 40$, we evaluate $\texttt{ctx} \in \{4, 8, 16, 32\}$ while keeping the remaining settings fixed, and we report the best checkpoint under the above validation criterion. For the reported multi-K Transformer runs, unless otherwise specified, we use $\texttt{ctx}= 16$, learning rate 3×10^{-5}, dropout 0.3, weight decay 0.1, batch size 256, and an application-level inner validation ratio of 0.2.

E. Evaluation Protocol

Given a test trace converted into a role sequence, we create prediction instances at each position by using the observed prefix and predicting the next role. We report Recall@k for $k \in \{1, 3, 5, 10, 20\}$.

We report two evaluation modes: 1) *all edges*, which includes all transitions with at least one preceding role, and 2) the *trigram-subset*, which restricts evaluation to transitions with at least two preceding roles for fair comparison with trigram-based methods. Unless otherwise specified, all models are evaluated on the same set of trigram-subset transition edges for fair comparison.

We evaluate atypical transition detection by converting next-step prediction into a sampled-candidate ranking problem with negative alternatives. For each true transition, we form one positive candidate (the observed next role) and $N_{\text{neg}} = 10$ negative candidates by replacing the next role with roles sampled uniformly at random from the role vocabulary $\{1, \ldots, K\} \setminus \{c_{t+1}\}$. Special tokens such as `[PAD]` and `[BOS]` are excluded from this sampling. We use a fixed random seed for reproducibility. This protocol keeps the candidate set size fixed across examples and models, while providing a lightweight approximation of anomaly discrimination without requiring exhaustive scoring over all possible next roles. Because the negatives are sampled uniformly at random, the task measures discrimination against random incorrect alternatives rather than against hardest or semantically nearest negatives. All candidates are scored using the model's assigned probability for the next role under the same observed context. We report AUROC and Precision–Recall metrics (including Average Precision) to account for the resulting class imbalance.

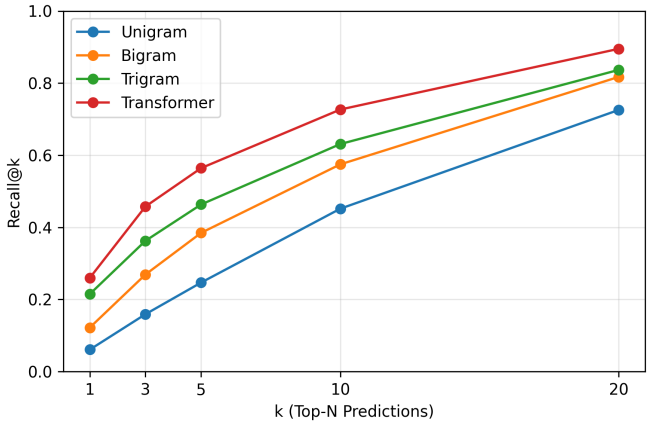


Fig. 3. Recall@k curves for next-step prediction on applications in the test split ($K = 40$), comparing unigram, bigram, trigram, and Transformer priors.

Accordingly, these metrics should be interpreted as measuring relative ranking performance under sampled alternatives, not as a direct estimate of real-world anomaly-detection accuracy under unconstrained deployment conditions.

VII. Results

We report results for the two tasks defined in Section III: 1) next-step suggestion via Recall@k and 2) atypical transition detection via ranking/classification metrics under negative sampling. Unless otherwise stated, results use the primary cross-app protocol (Section VI-C) and matched-edge evaluation for fair comparison.

A. Task 1: Next-Step Prediction Performance

Table III compares unigram, bigram, trigram, and the Transformer flow prior on the same matched set of 6,058 transition edges. The Transformer achieves the strongest Recall@k at every reported cutoff. Relative to the trigram baseline, the absolute improvements are +0.0449 at R@1, +0.0952 at R@3, +0.1009 at R@5, +0.0957 at R@10, and +0.0585 at R@20. Thus, the improvement is not simply monotonic in k; rather, the largest absolute gains occur in the shortlist range from R@3 to R@10, which is particularly relevant for shortlist-based design-assistance settings.

Fig. 3 visualizes the Recall@k curves and makes the differences in top-k ranking behavior easier to compare than the tabulated values alone.

Additional Transformer ablations over context length and regularization are reported in Appendix A.

To assess robustness to neural training randomness, we repeated the main $K = 40$ Transformer experiment with three random seeds while keeping the data split and model recipe fixed. As shown in Table IV, performance remains stable across seeds, with low standard deviation at all reported

TABLE II. KEY EXPERIMENTAL SETTINGS USED ACROSS MODELS

Component	Setting
Role vocabulary size	$K \in \{20, 40, 80, 120, 200\}$ (reported per experiment)
Duplicate collapsing	collapse consecutive identical roles; drop traces < 2
n-gram backoff	trigram–bigram interpolation (Eq. 6)
n-gram λ	tuned on validation split (maximizing Recall@10 on matched subset)
Transformer architecture	causal/autoregressive Transformer with masked self-attention
Transformer width/depth	$d = 256$, 4 heads, 4 layers
Transformer objective	cross-entropy next-role prediction
Transformer optimizer	AdamW
Transformer context length	ctx $\in \{4, 8, 16, 32\}$ (controlled sweep at $K = 40$)
Model selection	early stop on validation Recall@10 (trigram-subset), best checkpoint chosen
Training epochs	up to 200 with early stopping
Gradient clipping	max norm $= 1.0$
Transformer reported multi-K recipe	ctx $= 16$, lr $= 3 \times 10^{-5}$, dropout $= 0.3$, weight decay $= 0.1$, batch size $= 256$
Random seed	42 for the primary reported run; additional multi-seed robustness results are reported in Appendix E
Anomaly protocol	primary setting uses $N_{\text{neg}} = 10$; see App. F and App. G

TABLE III. RECALL@K FOR NEXT-STEP PREDICTION

Model	R@1	R@3	R@5	R@10	R@20
Unigram	0.0609	0.1591	0.2469	0.4518	0.7257
Bigram	0.1220	0.2692	0.3848	0.5748	0.8173
Trigram	0.2149	0.3627	0.4635	0.6311	0.8367
Transformer	**0.2598**	**0.4579**	**0.5644**	**0.7268**	**0.8952**

Recall@k for next-step prediction on unseen test applications with $K = 40$. All models are evaluated on the same 6,058 matched transition edges for fair comparison.

cutoffs: 0.2671 ± 0.0066 at R@1, 0.4691 ± 0.0097 at R@3, 0.5817 ± 0.0152 at R@5, 0.7436 ± 0.0146 at R@10, and 0.9032 ± 0.0071 at R@20. This indicates that the Transformer advantage is not tied to a favorable initialization, but persists across independent runs. Seed-specific results are provided in Appendix E.

1) Modality ablation at $K = 40$: To assess the contribution of each screen-representation modality, we repeated the $K = 40$ pipeline using three alternatives for screen-role induction: CLIP-only visual embeddings, SBERT-only hierarchy-text embeddings, and the fused CLIP+SBERT representation used in the main experiments. For each representation, we re-ran clustering, trace-to-role conversion, and downstream next-step modeling under the same application-level split and evaluation protocol. Table V shows that no single modality dominates uniformly across all models and recall cutoffs. For the Transformer, the fused representation yields the strongest top-ranked performance at $R@5$ and is effectively tied with SBERT-only at $R@1$, while CLIP-only slightly outperforms fused at $R@10$. This pattern suggests that fusion is best interpreted as a balanced representation choice rather than as a universally superior modality under all retrieval cutoffs.

We use $K = 40$ as the primary operating point in the remainder of this section. This choice is not based on raw next-step recall alone, since smaller values of K make the prediction problem easier by construction. Instead, $K = 40$ was selected in Section IV-C using clustering stability and fragmentation as the primary criteria, with downstream prediction used as a secondary check that the resulting role vocabulary remains predictive on held-out applications. Additional low-granularity diagnostics at $K = 10$ and $K = 15$ show that, although smaller values yield higher raw Recall@k, they also produce substantially coarser role vocabularies with much larger cluster sizes. We therefore use $K = 20$ as the coarse-grained lower bound in the downstream sweep and $K = 40$ as the representative operating point because it provides finer role granularity while still maintaining strong downstream performance.

Table VI reports next-step prediction performance across cluster granularities $K \in \{20, 40, 80, 120, 200\}$ under the same unseen test-application evaluation protocol; additional low-granularity results for $K = 10$ and $K = 15$ are reported in the Appendix. At $K = 20$, the $R@20$ column is saturated (1.0000 for all models) because the evaluation cutoff equals the role vocabulary size. This is a limitation of the current evaluation protocol at this granularity: with a coarse role space, $R@20$ no longer provides discriminative evidence about model quality and instead becomes a trivial upper bound. In other words, the task at this granularity is too coarse for $R@20$ to be informative under the current setup. Accordingly, the meaningful comparisons at $K = 20$ are $R@1$, $R@3$, $R@5$, and $R@10$.

As expected, raw Recall@k decreases as K increases, because the prediction task becomes harder when the role vocabulary is larger and the transition distribution becomes sparser. Across all K, the Transformer consistently outperforms the n-gram baselines.

To test out-of-domain generalization beyond application identity, we additionally evaluate leave-one-category-out with $K = 40$. Macro-averaged results over 26 held-out categories are shown in Table VII. The Transformer achieves the best macro-averaged Recall@k across all reported cutoff values, indicating stronger average category-level generalization than the n-gram baselines.

TABLE IV. MULTI-SEED TRANSFORMER NEXT-STEP PREDICTION RESULTS ($K = 40$)

Model	R@1	R@3	R@5	R@10	R@20
Transformer (3 seeds)	0.2671 ± 0.0066	0.4691 ± 0.0097	0.5817 ± 0.0152	0.7436 ± 0.0146	0.9032 ± 0.0071

Results are reported as mean $\pm$ standard deviation over three random seeds on unseen test applications.

TABLE V. MODALITY ABLATION ACROSS NEXT-STEP MODELS AT $K = 40$

Representation	Model	$R@1$	$R@5$	$R@10$
CLIP-only	Unigram	0.0668	0.2974	0.4797
Fused	Unigram	0.0950	0.3158	0.4958
SBERT-only	Unigram	0.0824	0.3042	0.4586
CLIP-only	Bigram	0.1198	0.4031	0.6127
Fused	Bigram	0.1455	0.4227	0.6104
SBERT-only	Bigram	0.1403	0.3946	0.5704
CLIP-only	Trigram	0.1967	0.4659	0.6448
Fused	Trigram	0.2199	0.4714	0.6367
SBERT-only	Trigram	0.2245	0.4750	0.6267
CLIP-only	Transformer	0.2224	0.5303	**0.7300**
Fused	Transformer	**0.2465**	**0.5629**	0.7216
SBERT-only	Transformer	0.2464	0.5525	0.6908

Each representation is used to induce screen roles, after which the same downstream application-level split and evaluation protocol are applied.

B. Task 2: Atypical Transition Detection

We evaluate atypical transition detection by ranking the true next role against $N_{\text{neg}} = 10$ randomly sampled incorrect roles per transition (Section VI-E). On the unseen test set, the Transformer achieves strong ranking performance, with AUROC = 0.83 and Average Precision (AP) = 0.42 under this protocol. These values indicate how well the model ranks observed transitions against sampled random alternatives under a fixed candidate set, and should not be interpreted as direct evidence of real-world anomaly-detection performance.

We compare n-gram baselines and the Transformer under an identical evaluation protocol: The same 6,058 test transitions, fixed negative sampling with $N_{\text{neg}} = 10$ (positive rate ≈ 0.0909), and the same sampled negatives reused across models. Table VIII reports AUROC, Average Precision (AP), and recall at a precision-constrained operating point (precision ≥ 0.8).

We also examined sensitivity to the number of sampled negatives in the anomaly protocol. Table IX shows that AUROC remains essentially unchanged across $N_{\text{neg}} \in \{5, 10, 20, 50\}$, varying only from 0.8317 to 0.8328. This suggests that the Transformer's relative ranking quality is robust to the size of the sampled candidate set. In contrast, Average Precision and recall at precision ≥ 0.8 decrease substantially as N_{neg} increases. This is expected, since larger N_{neg} makes the sampled task more imbalanced and pushes the high-precision operating point toward increasingly conservative thresholds. The result therefore strengthens the interpretation that the Transformer's anomaly advantage is robust as a ranking result, while threshold-dependent operating points remain sensitive to the difficulty of the sampled protocol. Additional results are reported in Appendix F.

To test whether the sampled-candidate protocol materially overstates anomaly-ranking performance, we additionally re-evaluated the main Transformer checkpoint against the full role vocabulary at $K = 40$ on the same 6,058 matched test transitions. Under this stricter setting, AUROC remains essentially unchanged (0.8320, versus 0.8323 under sampled negatives with $N_{\text{neg}} = 10$), while Average Precision decreases from 0.4231 to 0.1943 and recall at precision ≥ 0.8 decreases from 0.0568 to 0.0028. This indicates that the Transformer's relative ranking quality is robust beyond the sampled-negative setting, but that threshold-dependent anomaly screening becomes substantially more conservative when each context is evaluated against the full vocabulary.

To examine threshold behavior under different precision constraints, Table X reports thresholds selected to maximize recall while maintaining a target precision level. The strictest settings (e.g., target precision 0.95 or 0.90) achieve extremely low recall, indicating that these thresholds should be interpreted as highly conservative operating points rather than broad anomaly-screening regimes.

These results show that, although the Transformer provides the strongest overall ranking performance under the sampled evaluation protocol, very high precision comes at a substantial cost in recall. At target precision 0.95 and 0.90, only a very small fraction of atypical transitions is recovered. Lower precision targets such as 0.70, 0.60, or 0.50 provide substantially higher recall, but this comes at the expected cost of admitting more false positives.

Accordingly, Table X should be interpreted as illustrating the precision–recall trade-off under the fixed negative-sampling protocol rather than as establishing a deployment threshold for real-world anomaly detection.

Fig. 4 visualizes the full precision–recall curve and highlights the operating points in Table X.

C. Qualitative Analysis of Detected Redirect Failures

To complement quantitative ranking metrics, we report qualitative examples of high-confidence anomalies detected by the Transformer. Examples are selected from cases in which the observed transition is assigned low probability and there is a large margin between its probability and that of the top predicted continuation.

Fig. 5 shows an example from the KAYAK application. After a system permission request, the observed transition redirects the user to a hotel sharing interface that is unrelated to the permission context. The model assigns low probability to the observed continuation ($p_{\text{obs}} \approx 0.0035$) while ranking an account login/linking screen as the most likely continuation ($p_{\text{exp}} \approx 0.34$).

Fig. 6 shows an example from a transportation-related application. From a tariff configuration screen, the observed transition navigates to a driver configuration page without an apparent authentication step. The model instead predicts a login screen as the most likely continuation, assigning substantially higher probability than the observed transition.

TABLE VI. NEXT-STEP PREDICTION PERFORMANCE ACROSS CLUSTER GRANULARITIES

K	Model	R@1	R@3	R@5	R@10	R@20
20	Unigram	0.1176	0.2837	0.4146	0.6978	**1.0000**
20	Bigram	0.1588	0.3606	0.5083	0.7589	**1.0000**
20	Trigram	0.2851	0.4570	0.5830	0.7915	**1.0000**
20	Transformer	**0.3237**	**0.5442**	**0.6528**	**0.8365**	**1.0000**
40	Unigram	0.0609	0.1591	0.2469	0.4518	0.7257
40	Bigram	0.1220	0.2692	0.3848	0.5748	0.8173
40	Trigram	0.2149	0.3627	0.4635	0.6311	0.8367
40	Transformer	**0.2598**	**0.4579**	**0.5644**	**0.7268**	**0.8952**
80	Unigram	0.0440	0.1187	0.1690	0.2897	0.4743
80	Bigram	0.0993	0.2175	0.3015	0.4501	0.6365
80	Trigram	0.1567	0.2863	0.3638	0.4933	0.6655
80	Transformer	**0.2433**	**0.4023**	**0.4928**	**0.6284**	**0.7757**
120	Unigram	0.0337	0.0860	0.1273	0.2144	0.3632
120	Bigram	0.0876	0.1933	0.2648	0.3892	0.5416
120	Trigram	0.1217	0.2359	0.3094	0.4193	0.5685
120	Transformer	**0.2030**	**0.3528**	**0.4360**	**0.5588**	**0.6953**
200	Unigram	0.0171	0.0493	0.0770	0.1384	0.2493
200	Bigram	0.0734	0.1580	0.2203	0.3156	0.4436
200	Trigram	0.0959	0.1882	0.2497	0.3416	0.4621
200	Transformer	**0.1783**	**0.3095**	**0.3811**	**0.4956**	**0.6107**

Next-step prediction performance across cluster granularities. All results are evaluated on unseen test applications using cumulative Recall@k.

TABLE VII. LEAVE-ONE-CATEGORY-OUT RESULTS

Model	R@1	R@3	R@5	R@10
Unigram	0.0640	0.1604	0.2637	0.4631
Bigram	0.1217	0.2770	0.3891	0.5693
Trigram	0.1963	0.3590	0.4596	0.6283
Transformer	**0.2105**	**0.4115**	**0.5257**	**0.6985**

Leave-one-category-out

macro-average results over 26 held-out categories ($K = 40$), evaluated on the trigram-subset.

TABLE VIII. ATYPICAL TRANSITION DETECTION

Model	AUROC ↑	AP ↑	Recall @ Precision ≥ 0.8 ↑
Unigram	0.665	0.149	0.0000
Bigram	0.746	0.242	0.0000
Trigram	0.771	0.336	0.0154
Transformer	0.832	0.423	0.0568

Atypical

transition detection under a fixed protocol (same 6,058 transitions; fixed negative sampling with $N_{\text{neg}} = 10$; positive rate ≈ 0.0909). We report AUROC, Average Precision (AP), and recall at precision ≥ 0.8.

TABLE IX. SENSITIVITY OF TRANSFORMER ATYPICAL-TRANSITION METRICS TO THE NUMBER OF SAMPLED NEGATIVES N_{neg} ($K = 40$)

N_{neg}	AUROC ↑	AP ↑	Recall@Precision ≥ 0.8 ↑
5	0.8317	0.5626	0.1720
10	0.8323	0.4231	0.0568
20	0.8319	0.2957	0.0102
50	0.8328	0.1654	0.0012

TABLE X. PRECISION-RECALL TRADE-OFFS FOR TRANSFORMER-BASED ATYPICAL TRANSITION DETECTION

Target Precision	Threshold	Achieved Precision	Recall
0.95	0.9114	1.0000	0.00033
0.90	0.8618	0.9231	0.00396
0.70	0.3774	0.7003	0.1354
0.60	0.1904	0.6000	0.2565
0.50	0.1030	0.5000	0.3760

Thresholds are chosen to maximize recall while satisfying a target precision constraint. Very high target precision yields highly conservative operating points with limited recall.

VIII. DISCUSSION

Our results support the central hypothesis of this work: after abstracting screens into reusable role IDs, mobile UI navigation exhibits domain-independent regularities that can be learned from large-scale traces and transferred to unseen applications and categories. We discuss implications for next-step suggestion and atypical transition detection, analyze the effect of role granularity, and outline limitations and future directions.

A. Next-Step Prediction on Unseen Applications

Across unseen test applications, the Transformer flow prior consistently outperforms count-based n-gram baselines, with the largest absolute gains occurring in the shortlist range from R@3 to R@10 (Table III, Fig. 3). For a fixed shortlist size, this means that the Transformer more often places the true next role among the top-ranked candidates, which is compatible with shortlist-based design-assistance settings in which a small set of plausible continuations is presented to the user.

A plausible interpretation is that role sequences contain recurring multi-step patterns (e.g., transitions from entry to browsing to detail to confirmation states) that benefit from sequence models able to condition on longer histories, whereas n-gram baselines are limited to fixed-order local context. However, the present experiments establish the empirical advantage of the Transformer rather than the exact mechanism responsible for it.

The sensitivity study across K (Table VI) further clarifies the role of abstraction. As K increases, the next-role classification problem becomes harder because the role vocabulary is larger and the transition distribution becomes sparser. This trend affects all models, but the degradation is steeper for frequency-based baselines that rely directly on repeated transition counts. In contrast, the Transformer maintains stronger performance across the full range of tested granularities, indicating that it is better able to preserve useful ranking quality as the role abstraction becomes finer.

Throughout the main results we report $K = 40$ as the primary operating point. This choice is not based on raw next-step recall alone, since smaller values of K make the prediction task easier by construction. Instead, $K = 40$ was selected as a practical trade-off between representational granularity,

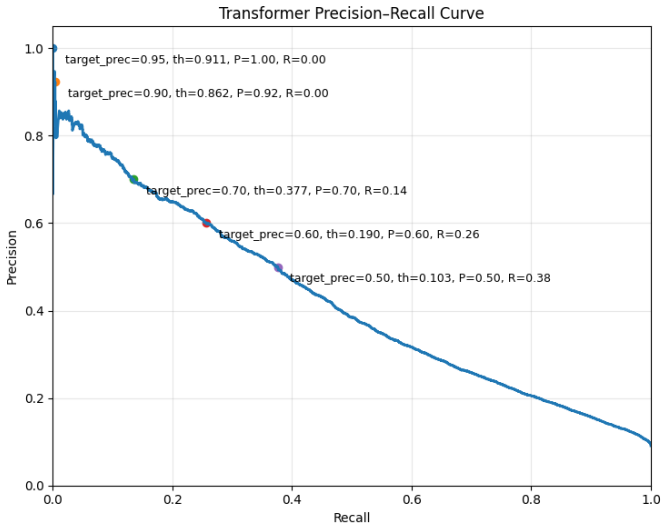


Fig. 4. Precision-recall curve for Transformer-based atypical transition detection under negative sampling. Markers correspond to selected precision-constrained operating points.

clustering stability, and downstream predictive performance, supported by the diagnostics in Section IV-C together with the sensitivity analysis in Table VI. We nevertheless report results across multiple values of K and treat role granularity as a tunable design parameter rather than a fixed constant.

B. Generalization Beyond Applications: Cross-Category Behavior

The leave-one-category-out protocol tests a more challenging distribution shift than cross-application splits, since the held-out category may involve different UI conventions, vocabulary, and navigational patterns. The n-gram results (Table VII) show that the role abstraction retains predictive value even under this stronger shift. The Transformer further improves macro-averaged Recall@k under the same protocol, providing evidence of stronger average generalization across held-out categories than the count-based baselines.

At the same time, these results should be interpreted conservatively. Categories in RICO are not perfectly disjoint in UI structure, and some categories may still share common design conventions such as list-detail patterns or authentication flows. The full per-category breakdown confirms that the macro-average masks meaningful heterogeneity rather than uniform behavior across held-out categories. In the present experiments, Transformer R@10 ranges from 0.5446 on *Sports* and 0.5843 on *Food & Drink* to 0.7666 on *Education*, 0.7455 on *Weather*, and 0.7452 on *Beauty*. Thus, the leave-one-category-out result should be read as evidence of strong average cross-category generalization together with substantial variation in category difficulty. Full per-category results are reported in Appendix H.

A plausible explanation for the weaker held-out results on *Sports* and *Food & Drink* is that these categories often involve more heterogeneous and content-driven flows than the strongest held-out categories. Sports applications may combine news, live scores, video, team pages, betting- or fantasy-like features, and account flows within the same app family, while Food & Drink applications may mix restaurant discovery, menus, ordering, reservation, delivery, and loyalty flows. Such mixtures can produce more branching continuations from visually similar intermediate states and therefore reduce next-role predictability under a shared role vocabulary. By contrast, stronger categories such as Education, Weather, and Beauty may rely more heavily on recurring template-like flows such as onboarding, settings, browsing, and detail views, which transfer more readily across applications. These explanations are interpretive rather than directly annotated in the present dataset, but they are consistent with the heterogeneity visible in the full leave-one-category-out breakdown.

C. Interpreting Atypical Transition Detection Metrics

The atypical-transition results in this study should be interpreted within the scope of the evaluation protocol. Specifically, the task is defined through fixed negative sampling, so the reported metrics measure how well each model ranks observed transitions against sampled alternatives under this controlled setting. They therefore support comparison of relative ranking performance across models, but do not by themselves establish real-world anomaly-detection effectiveness under unconstrained deployment conditions.

In particular, the protocol does not test whether low-probability transitions correspond to naturally occurring defects in deployment, since sampled negatives may differ substantially from realistic alternative continuations.

At the same time, an additional full-vocabulary re-evaluation at $K = 40$ shows that AUROC remains nearly unchanged relative to the sampled protocol, whereas Average Precision and precision-constrained recall drop substantially, indicating that sampled negatives preserve relative ranking comparisons reasonably well but understate the difficulty of strict high-precision anomaly screening. Even with this stricter full-vocabulary comparison, the evaluation remains an offline atypical-transition ranking task rather than a direct benchmark of naturally occurring deployment anomalies.

Within this fixed protocol, the Transformer achieves the strongest ranking performance among the compared models (Table VIII), indicating better discrimination between observed transitions and sampled negative alternatives than the n-gram baselines. Relative to the trigram baseline, this result suggests that atypical-transition scoring may benefit from information beyond short local context, although the present experiments do not directly establish the underlying mechanism.

At the same time, the precision-constrained analysis (Table X, Fig. 4) highlights an important limitation. At very high precision targets such as 0.95 or 0.90, recall becomes extremely low. These operating points should therefore not be interpreted as evidence of broad practical usefulness; rather, they show that only a very small fraction of positives can be recovered when false alarms are tightly constrained. Lower precision targets provide substantially higher recall, but this comes at the expected cost of admitting more false positives.

Overall, the anomaly results are best understood as evidence that the Transformer provides the strongest relative ranking performance under the sampled evaluation protocol, while also revealing clear practical limitations in strict high-precision regimes. A more realistic assessment of deployment

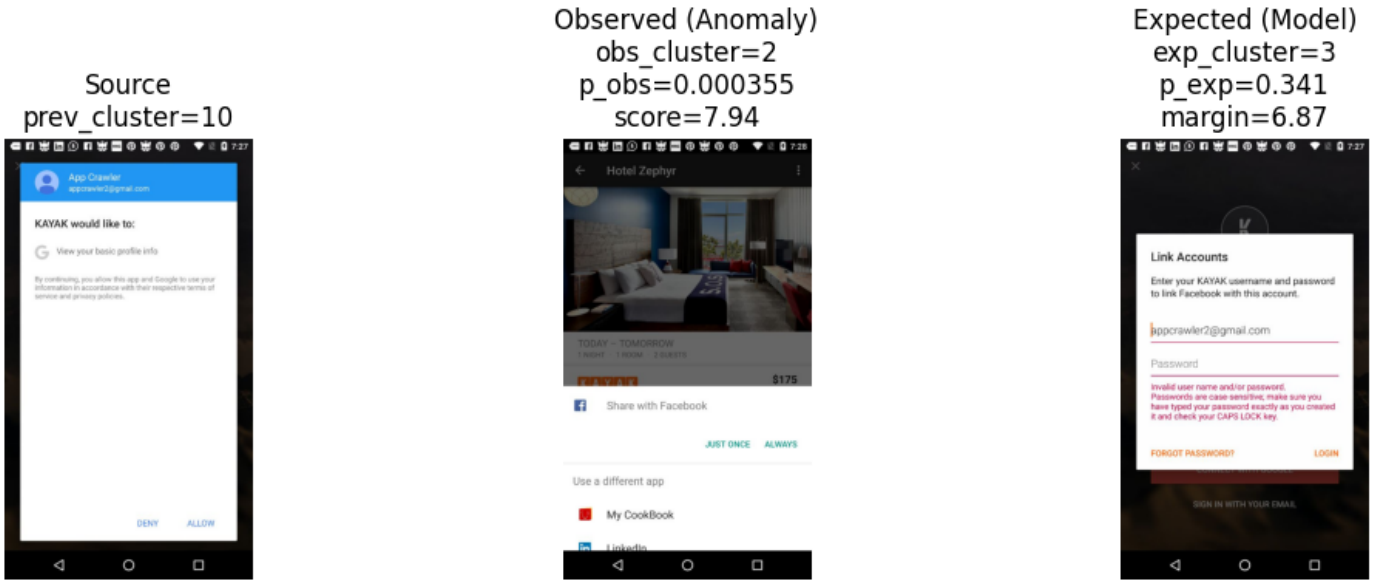


Fig. 5. Detected redirect failure in the KAYAK application. After a permission request (left), the observed transition leads to an unrelated sharing interface (middle), while the model predicts an account login screen as the expected continuation (right).



Fig. 6. Detected configuration flow anomaly. From a tariff configuration screen (left), the application navigates to a driver settings page without authentication (middle). The model instead predicts a login screen as the expected continuation (right).

utility would require evaluation settings that go beyond sampled negatives and better reflect naturally occurring atypical transitions.

D. Qualitative Inspection of High-Scoring Transitions

The qualitative case studies (Fig. 5 and 6) are interpreted as illustrative examples rather than as a general validation of the anomaly score. In the presented cases, high anomaly scores coincide with transitions that appear questionable or unusual from a design perspective, while the model assigns substantially higher probability to an alternative continuation that is more consistent with the preceding flow context (e.g., an authentication or account-linking step). These examples therefore suggest that the score can highlight potentially design-relevant irregularities in at least some cases, but they do not establish that this holds broadly beyond the illustrated examples.

The contrast between the observed continuation and the model-preferred alternative can nevertheless provide a qualitative basis for inspection. A designer can examine both what actually occurred and what the model ranked as more probable under the learned flow prior. This comparison may

help interpret why a transition was flagged, although the present qualitative evidence is not sufficient to claim general diagnostic usefulness.

A limitation is that qualitative selection can be biased toward visually obvious failures. To mitigate this, future qualitative evaluation should include a structured sampling strategy (e.g., top anomalies stratified by application/category, plus a random sample of mid-score transitions) together with human annotation to estimate how often low-probability transitions correspond to true flow defects versus legitimate but rare paths.

E. Role Granularity and the Choice of K

The experiments do not support selecting K from next-step prediction accuracy alone. Smaller values of K produce higher Recall@k, but also correspond to coarser role vocabularies in which more heterogeneous screens are merged together. Conversely, larger values of K preserve finer distinctions but lead to increased sparsity and weaker transition modeling performance.

On this basis, we use $K = 40$ as the main operating point in the study. This choice is supported by two considerations grounded in the reported experiments: 1) downstream predictive behavior on held-out applications, where $K = 40$ retains strong next-step prediction performance without relying on the overly coarse abstraction induced by smaller K values; and 2) clustering diagnostics in Section IV-C, where $K = 40$ shows favorable stability while avoiding the stronger fragmentation observed at larger values of K. We therefore treat $K = 40$ not as a universally "natural" constant, but as a practically selected operating point supported by the present evidence.

More generally, the experiments suggest that role granularity should be treated as a tunable design parameter whose preferred value depends on the balance one wants between semantic specificity and statistical reliability.

F. Limitations and Future Work

This study has several limitations. First, the primary atypical-transition evaluation uses negative sampling, which provides a controlled quantitative protocol but does not fully capture real-world anomalies. We complement this with a full-vocabulary re-evaluation at $K = 40$, but both settings remain offline ranking evaluations rather than direct measurements of naturally occurring deployment failures. In practice, unusual transitions may depend on latent application or user state, such as authentication status, granted permissions, network conditions, or prior interaction history, none of which is explicitly modeled in the current setup. As a result, the anomaly results should be interpreted as evidence of ranking quality under controlled candidate-comparison settings rather than as a complete evaluation of real deployment conditions.

Second, we added four robustness analyses to reduce dependence on a single reported configuration: multi-seed evaluation for the Transformer, sensitivity analysis with respect to the number of sampled negatives N_{neg}, an exhaustive full-vocabulary evaluation at $K = 40$, and a full per-category breakdown under leave-one-category-out. Together, these analyses show that the main next-step gains are stable across random seeds, that anomaly-ranking performance remains stable in AUROC across different sampled-negative settings, that exhaustive comparison is substantially harsher than sampled evaluation while still preserving useful ranking quality, and that cross-category generalization remains strong on average despite substantial heterogeneity across held-out categories. The remaining limitation is therefore not the absence of these checks, but the fact that anomaly evaluation is still defined primarily through sampled negatives rather than naturally occurring deployment anomalies. An additional full-vocabulary re-evaluation at $K = 40$ shows that AUROC is largely unchanged relative to the sampled protocol, but that Average Precision and precision-constrained recall decrease substantially, so the sampled setting appears to preserve relative ranking comparisons better than it preserves realistic high-precision screening difficulty.

Beyond these limitations, several research extensions are promising. One direction is to design more realistic anomaly benchmarks that better account for latent state, for example by constructing negatives within matched application/session contexts or by incorporating human-labeled subsets of truly atypical transitions. Another is to expand the comparison set beyond n-gram baselines and a causal Transformer, for example by evaluating recurrent models, structured probabilistic sequence models, or retrieval-augmented priors, in order to better understand which sequence-modeling assumptions are most suitable for UI flow data. Finally, because the current role abstraction is unsupervised and intentionally label-free, an important direction is to generate lightweight role descriptors (e.g., by summarizing common hierarchy tokens or nearest-neighbor exemplars) to improve interpretability and debugging without requiring manual role annotation. While this qualitative audit improves interpretability, it does not replace a formal human evaluation of cluster semantics. A larger-scale annotated study remains an important direction for future work.

IX. Conclusion

We presented a role-based formulation of mobile UI navigation in which screens are mapped to discrete, unsupervised *screen roles* using multimodal representations and clustering, allowing raw interaction traces to be modeled as role sequences across applications. This abstraction supports cross-application sequence modeling while preserving enough structure to study downstream navigation behavior.

Across unseen test applications, both n-gram baselines and a causal Transformer capture nontrivial regularities in role transitions, with the Transformer consistently achieving the strongest next-step prediction performance. Under the fixed negative-sampling protocol used for atypical-transition evaluation, the Transformer also achieves the strongest ranking performance among the compared models. Sensitivity analysis across role granularities further shows that the choice of K should be treated as a practical design parameter rather than a fixed constant; in this study, $K = 40$ serves as a representative operating point supported by clustering diagnostics and downstream predictive behavior.

The qualitative examples illustrate that some high-scoring atypical transitions can correspond to potentially design-relevant irregularities, although broader validation would require more realistic anomaly benchmarks and structured hu-

man evaluation. More generally, the results indicate that role-based abstractions provide a promising basis for studying UI flow regularities and for ranking-oriented tasks such as next-step suggestion under cross-application generalization.

The robustness analyses further indicate that these conclusions are stable across random seeds, across multiple negative-sampling settings, under a stricter full-vocabulary comparison, and across detailed held-out-category breakdowns. Future work should therefore focus less on basic stability verification and more on richer latent-state information, improved interpretability through lightweight role descriptors, and more realistic anomaly-evaluation protocols.

ACKNOWLEDGMENT

The authors would like to thank the creators of the *Rico* dataset for making it publicly available, which made this research possible.

CODE AVAILABILITY

The source code implementing the models and training procedures used in this study is available at: https://github.com/salsik/trico. The repository also includes the exact application-level split manifests used in the main experiments, including the train/validation/test app partitions for the cross-application protocol and the held-out category lists for the leave-one-category-out evaluation.

REFERENCES

[1] B. Deka, Z. Huang, C. Franzen, J. Hibschman, D. Afergan, Y. Li, J. Nichols, and R. Kumar, "Rico: A mobile app dataset for building data-driven design applications," in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*, 2017, pp. 845–854.

[2] T. J.-J. Li, L. Popowski, T. M. Mitchell, and B. A. Myers, "Screen2vec: Semantic embedding of gui screens and gui components," *arXiv preprint arXiv:2101.11103*, 2021. [Online]. Available: https://arxiv.org/abs/2101.11103

[3] B. Wang, G. Li, X. Zhou, Z. Chen, T. Grossman, and Y. Li, "Screen2words: Automatic mobile ui summarization with multimodal learning," in *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21)*, 2021.

[4] L. A. Leiva, A. Hota, and A. Oulasvirta, "Enrico: A dataset for topic modeling of mobile ui designs," in *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20 Extended Abstracts)*, 2020.

[5] L. Fan, T. Su, S. Chen, G. Meng, Y. Liu, and L. Xu, "Storydroid: Automated generation of storyboard for android apps," in *Proceedings of the 41st International Conference on Software Engineering (ICSE '19)*, 2019.

[6] X. Zhou and Y. Li, "Large-scale modeling of mobile user click behaviors using deep learning," in *Proceedings of the 15th ACM Conference on Recommender Systems (RecSys '21)*, 2021.

[7] S. Feiz, J. Wu, X. Zhang, A. Swearngin, T. Barik, and J. Nichols, "Understanding screen relationships from screenshots of smartphone applications," in *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*, 2022.

[8] B. Deka, Z. Huang, and R. Kumar, "Erica: Interaction mining mobile apps," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*, 2016, pp. 767–776.

[9] G. Hu, L. Zhu, and J. Yang, "Appflow: Using machine learning to synthesize robust, reusable ui tests," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '18)*, 2018.

[10] J. Wu, Y.-H. Peng, X. Y. A. Li, A. Swearngin, J. P. Bigham, and J. Nichols, "Uiclip: A data-driven model for assessing user interface design," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*, 2024, pp. 45:1–45:16.

[11] I. Salama, L. H. Mormille, and M. Atsumi, "Laycoder: UI layout completion with an encoder-only transformer and layout tokenizer," *International Journal of Advanced Computer Science and Applications*, vol. 17, no. 2, 2026.

[12] T. J.-J. Li, A. Azaria, B. A. Myers, and J. Nichols, "Kite: Building conversational bots from mobile apps," in *Proceedings of the 16th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys '18)*, 2018.

[13] G. Li, X. Zhang, A. Swearngin, and J. Nichols, "Sift: Grounding multimodal language models for interactive ui tasks," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2023.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of EMNLP-IJCNLP*, 2019.

[16] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th international conference on World Wide Web*, 2010, pp. 1177–1178.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: https://www.jmlr.org/papers/v12/pedregosa11a.html

[18] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[19] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 224–227, 1979.

[20] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[21] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, Dec. 2002. [Online]. Available: https://www.jmlr.org/papers/v3/strehl02a.html

[22] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.

[23] D. Jurafsky and J. H. Martin, "Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models," 2026, online manuscript released January 6, 2026. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

APPENDIX

This appendix provides additional experimental details and supplementary results that support the main work.

A. Transformer Ablations for Next-Step Prediction

We conducted a broad hyperparameter sweep for Transformer-based next-step prediction under the cross-application protocol, using $K = 40$ as the main operating point. In total, the sweep comprised 120 training runs. The goal of this search was to verify that the reported Transformer

results are stable under reasonable variations in sequence context and training hyperparameters, and to select the checkpoint used in the main experiments.

Across the sweep, we varied maximum context length, learning rate, dropout, and the application-level inner validation ratio. Specifically, the explored values were $\texttt{ctx} \in \{4, 8, 16, 32\}$, learning rate $\in \{3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}\}$, dropout $\in \{0.1, 0.2, 0.3\}$, and validation ratio $\in \{0.05, 0.1, 0.2\}$. All runs used label smoothing 0.0, a maximum training budget of 200 epochs, and patience $= 8$. Table XI reports the strongest configurations from this 120-run sweep rather than listing all runs individually.

The best validation checkpoint used throughout the main experiments is obtained with context length 16, learning rate 3×10^{-5}, dropout 0.3, weight decay 0.1, batch size 256, and validation ratio 0.2, achieving a best validation score of 0.7187. Two general patterns are visible in the sweep. First, context length 16 consistently outperforms shorter context (8) and remains preferable to the longer setting (32), suggesting that moderate-length history is sufficient for most predictive gains in the role-sequence setting. Second, mild regularization improves selection performance, with dropout 0.3 and weight decay 0.1 appearing more reliably among the strongest runs than weaker regularization settings.

Because this selected checkpoint is the one used in all downstream experiments, the sweep serves both as a model-selection record and as evidence that the reported Transformer gains are not tied to a single arbitrary hyperparameter choice.

B. Transformer Implementation Details

The Transformer next-role model is implemented in PyTorch as a causal sequence model over discrete role IDs. Given a role vocabulary of size K, we reserve two additional token IDs for `[PAD]` and `[BOS]`, yielding a total token vocabulary of size $K + 2$. During training, each trace is converted into prefix–target pairs by prepending `[BOS]` and predicting each next role from its observed prefix.

Batch construction uses variable-length prefixes clipped to at most the selected context length `ctx`. Within each mini-batch, the effective sequence length is the minimum of `ctx` and the longest prefix in that batch. Prefixes longer than this limit are truncated to the most recent context, while shorter prefixes are right-padded with `[PAD]`. The attention mask marks valid tokens and is used both as a padding mask and to recover the last non-padding hidden state for prediction.

The model is implemented as a stack of Transformer encoder blocks equipped with a causal triangular mask, so that token representations depend only on the observed prefix despite using the encoder-layer primitive. The final hidden state at the last valid prefix position is passed through a linear output layer to obtain logits over the next-role vocabulary.

Training uses AdamW optimization, gradient clipping with maximum norm 1.0, shuffled mini-batches of size 256, and a fixed random seed of 42. The script allows up to 200 training epochs and selects the best checkpoint using early stopping on validation Recall@10 computed on the trigram-subset, with patience 8 and minimum improvement threshold 10^{-4}. The training, validation, and test partitions are all constructed at the application level.

The implementation also evaluates two edge sets for next-step prediction: *all edges*, which begin from the first prediction after `[BOS]`, and the *trigram-subset*, which includes only transitions with at least two preceding real roles in the observed sequence. The latter is used for checkpoint selection and for comparisons intended to align with higher-order context baselines.

The final checkpoint used in the main experiments corresponds to context length 16, learning rate 3×10^{-5}, dropout 0.3, weight decay 0.1, batch size 256, and validation ratio 0.2.

C. Additional Evaluation Details

We also evaluated a 4-gram baseline as a diagnostic of whether increasing Markov order improves atypical-transition ranking under the same fixed negative-sampling protocol. The 4-gram model did not outperform the trigram baseline (AUROC $= 0.752$, AP $= 0.309$), consistent with increased sparsity at higher orders.

For clarity, the atypical-transition results reported in the study are based on a fixed sampled discrimination setting rather than a fully natural anomaly benchmark. Accordingly, the reported AUROC, AP, and precision–recall values should be interpreted as measuring relative ranking quality under this controlled protocol.

D. Additional Low-Granularity Results for $K = 10$ and $K = 15$

To assess whether the downstream granularity sweep should extend below $K = 20$, we additionally evaluated $K = 10$ and $K = 15$. Table XII reports the corresponding next-step prediction results. These smaller values yield higher raw Recall@k, but correspond to substantially coarser role vocabularies, as reflected by their much larger median cluster sizes in Table I. For these low-granularity Transformer runs, checkpoint selection used validation Recall@5 rather than Recall@10, since Recall@10 becomes trivial at $K = 10$.

E. Multi-seed Transformer Results

Table XIII reports the seed-specific Transformer results underlying the summary in Table IV. These runs use the same $K = 40$ setting, data split, and model recipe as the main experiment, differing only in random seed. The low variation across seeds supports the stability of the reported Transformer gains.

F. Sensitivity to the Number of Sampled Negatives

We evaluated Transformer-based atypical-transition detection under four negative-sampling settings, $N_{\text{neg}} \in \{5, 10, 20, 50\}$. As shown in Table XIV, AUROC remains nearly constant across all settings (0.8317–0.8328), indicating stable relative ranking quality. In contrast, Average Precision decreases from 0.5626 at $N_{\text{neg}} = 5$ to 0.1654 at $N_{\text{neg}} = 50$, and recall at precision ≥ 0.8 decreases from 0.1720 to 0.0012. This behavior is expected because increasing N_{neg} makes the sampled evaluation more imbalanced and yields increasingly conservative high-precision thresholds. Table XIV additionally reports the selected decision threshold and the achieved precision for each setting.

TABLE XI. TOP-PERFORMING CONFIGURATIONS FROM TRANSFORMER SWEEP FOR NEXT-STEP PREDICTION AT $K = 40$

ctx	lr	dropout	val ratio	Best val. score
16	3×10^{-5}	0.3	0.2	**0.7187**
16	3×10^{-5}	0.3	0.1	0.7179
16	3×10^{-5}	0.2	0.2	0.7159
16	3×10^{-4}	0.3	0.2	0.7157
16	3×10^{-5}	0.1	0.2	0.7146
8	3×10^{-5}	0.3	0.2	0.7144
16	10^{-4}	0.3	0.2	0.7144
32	3×10^{-5}	0.3	0.2	0.7141

Top-performing configurations from the 120-run Transformer sweep for next-step prediction at $K = 40$ under the cross-application protocol. All runs used label smoothing 0.0, 200 training epochs, and patience $= 8$. The first row is the checkpoint used in the main experiments. For brevity, only the strongest configurations are shown.

TABLE XII. LOW-GRANULARITY NEXT-STEP PREDICTION RESULTS FOR $K = 10$ AND $K = 15$

K	Model	R@1	R@3	R@5	R@10
10	Unigram	0.1935	0.4698	0.6788	1.0000
10	Bigram	0.2472	0.5455	0.7780	1.0000
10	Trigram	0.3747	0.6476	0.8299	1.0000
10	Transformer	**0.4022**	**0.6976**	**0.8650**	**1.0000**
15	Unigram	0.1237	0.3672	0.5451	0.8835
15	Bigram	0.1732	0.4283	0.6196	0.9189
15	Trigram	0.3207	0.5519	0.7125	0.9305
15	Transformer	**0.3516**	**0.6193**	**0.7745**	**0.9440**

Additional low-granularity next-step prediction results used to assess whether the downstream granularity sweep should extend below $K = 20$. As expected, smaller values yield higher raw Recall@k, but correspond to substantially coarser role vocabularies. For the Transformer runs at $K = 10$ and $K = 15$, checkpoint selection used validation Recall@5 rather than Recall@10, since Recall@10 becomes trivial at $K = 10$.

TABLE XIII. SEED-SPECIFIC TRANSFORMER NEXT-STEP PREDICTION RESULTS FOR $K = 40$

Seed	R@1	R@3	R@5	R@10	R@20
42	0.2598	0.4579	0.5644	0.7268	0.8952
43	0.2729	0.4741	0.5928	0.7514	0.9054
44	0.2686	0.4754	0.5880	0.7527	0.9089
Mean $\pm$ Std	0.2671 ± 0.0066	0.4691 ± 0.0097	0.5817 ± 0.0152	0.7436 ± 0.0146	0.9032 ± 0.0071

TABLE XIV. SENSITIVITY OF TRANSFORMER ATYPICAL-TRANSITION METRICS TO THE NUMBER OF SAMPLED NEGATIVES N_{neg} ($K = 40$)

N_{neg}	AUROC ↑	AP ↑	Recall@Precision ≥ 0.8 ↑	Threshold	Achieved Precision
5	0.8317	0.5626	0.1720	0.3053	0.8003
10	0.8323	0.4231	0.0568	0.5935	0.8000
20	0.8319	0.2957	0.0102	0.8060	0.8158
50	0.8328	0.1654	0.0012	0.8986	0.8750

G. Full-Vocabulary Re-Evaluation

The full-vocabulary result in Table XV shows that the sampled-negative protocol does not materially inflate AUROC, but it does make the anomaly task appear less severe in terms of Average Precision and high-precision recall. This supports the interpretation that sampled negatives are acceptable for comparing relative ranking quality across models, while still understating the difficulty of precision-constrained screening under exhaustive candidate evaluation.

H. Full Leave-One-Category-Out Breakdown

Table XVI reports the full per-category leave-one-category-out results for the Transformer-based model at $K = 40$. The breakdown shows substantial heterogeneity across categories. The strongest held-out categories by Transformer R@10 are *Education* (0.7666), *Weather* (0.7455), and *Beauty* (0.7452), whereas the weakest are *Sports* (0.5446) and *Food & Drink* (0.5843). This supports the interpretation that the macro-average reported in the main text reflects strong average transfer, but not uniform difficulty across all UI domains.

I. Qualitative Audit of Representative Screen Roles

To assess whether the learned role abstraction at $K = 40$ corresponds to semantically coherent UI states rather than arbitrary visual partitions, we qualitatively inspected representative screens from selected high-frequency clusters. For each cluster, screens were ranked by proximity to the cluster centroid in the fused embedding space, and the nearest examples were retained while applying simple diversity constraints to reduce near-duplicate screens from the same app or trace.

Fig. 7–10 show representative examples for 12 selected clusters. These examples are intended as qualitative support

TABLE XV. COMPARISON BETWEEN THE PRIMARY SAMPLED-NEGATIVE ANOMALY PROTOCOL AND A FULL-VOCABULARY RE-EVALUATION FOR THE TRANSFORMER AT $K = 40$. BOTH SETTINGS USE THE SAME 6,058 MATCHED TEST TRANSITIONS

Evaluation setting	AUROC ↑	AP ↑	Recall@Precision ≥ 0.8 ↑
Sampled negatives ($N_{\text{neg}} = 10$)	0.8323	0.4231	0.0568
Full vocabulary (39 negatives/context)	0.8320	0.1943	0.0028

TABLE XVI. FULL LEAVE-ONE-CATEGORY-OUT BREAKDOWN FOR THE TRANSFORMER AT $K = 40$. RESULTS ARE REPORTED PER HELD-OUT CATEGORY ON THE TRIGRAM-SUBSET

Category	R@1	R@3	R@5	R@10
Education	0.2593	0.4856	0.6143	0.7666
Weather	0.1506	0.3933	0.5576	0.7455
Beauty	0.2258	0.3968	0.5290	0.7452
Dating	0.2272	0.4240	0.5786	0.7305
Video Players & Editors	0.2017	0.4177	0.5308	0.7282
Comics	0.2003	0.4035	0.5512	0.7281
House & Home	0.2778	0.4709	0.5450	0.7275
Entertainment	0.2176	0.4078	0.5285	0.7255
Auto & Vehicles	0.2320	0.4394	0.5524	0.7248
Art & Design	0.2238	0.4476	0.5429	0.7190
Business	0.2206	0.4428	0.5492	0.7162
Parenting	0.2186	0.4269	0.5439	0.7126
Social	0.2158	0.4200	0.5406	0.7119
Finance	0.2336	0.4252	0.5245	0.7023
Communication	0.1822	0.3779	0.5187	0.7006
Medical	0.2312	0.4363	0.5462	0.6946
Books & Reference	0.1973	0.3867	0.5050	0.6938
Lifestyle	0.1955	0.3849	0.5045	0.6905
Health & Fitness	0.2337	0.4376	0.5290	0.6861
Shopping	0.1952	0.3921	0.5064	0.6745
Maps & Navigation	0.2241	0.4212	0.5140	0.6742
Music & Audio	0.2406	0.4230	0.5235	0.6682
News & Magazines	0.2018	0.3691	0.4877	0.6587
Travel & Local	0.1834	0.3846	0.5022	0.6575
Food & Drink	0.1776	0.3582	0.4359	0.5843
Sports	0.1392	0.2948	0.3835	0.5446

for the interpretability of the learned role abstraction, complementing the internal clustering diagnostics reported in the main study. They should not be interpreted as a substitute for a formal human-annotation study.

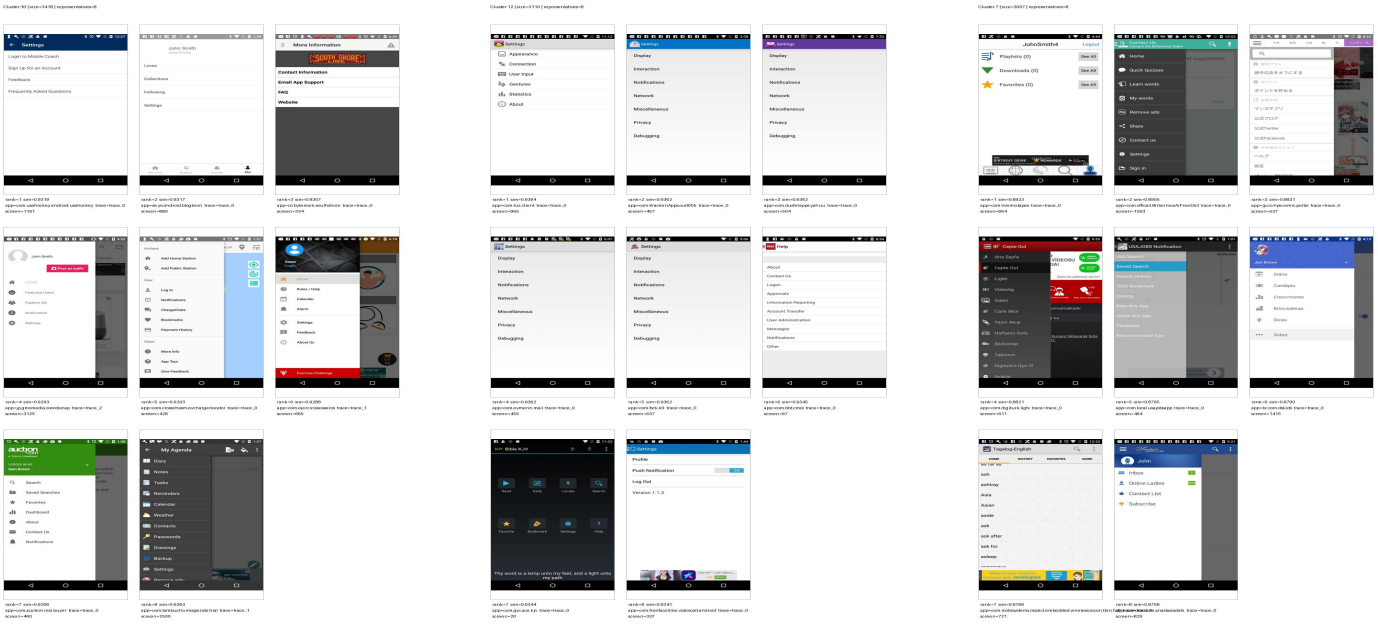


Fig. 7. Qualitative audit of representative learned screen roles at $K = 40$ (Part I).

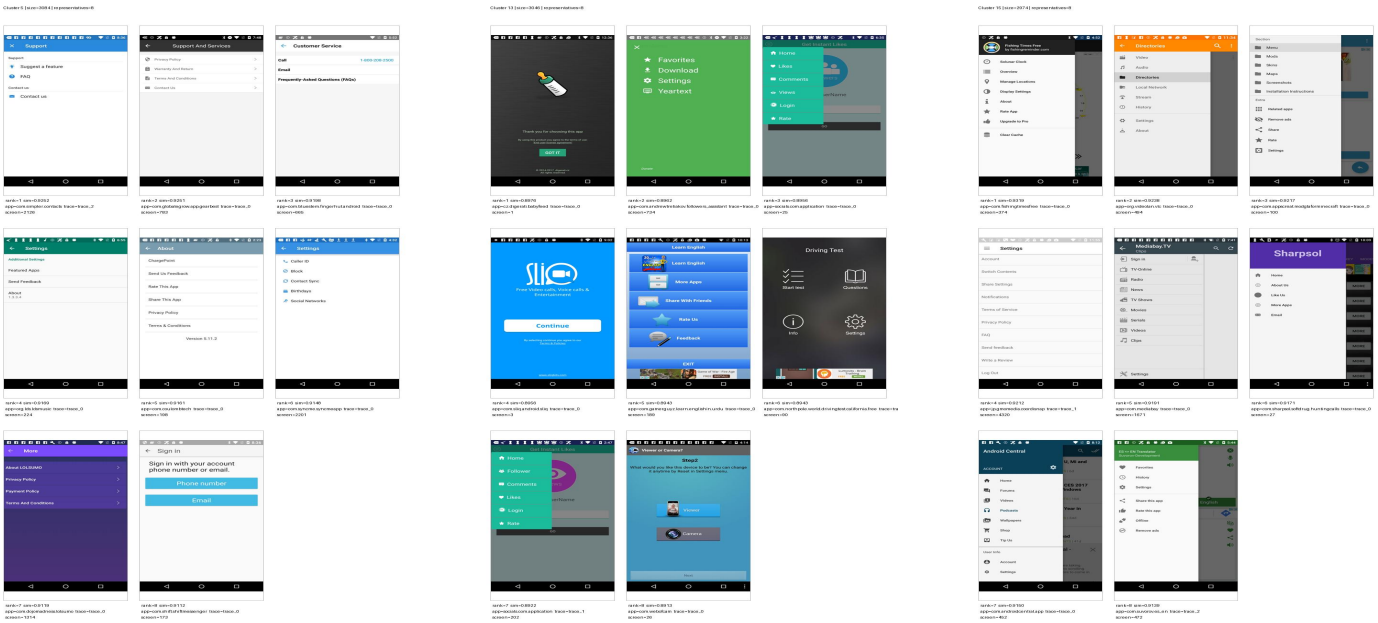


Fig. 8. Qualitative audit of representative learned screen roles at $K = 40$ (Part II).

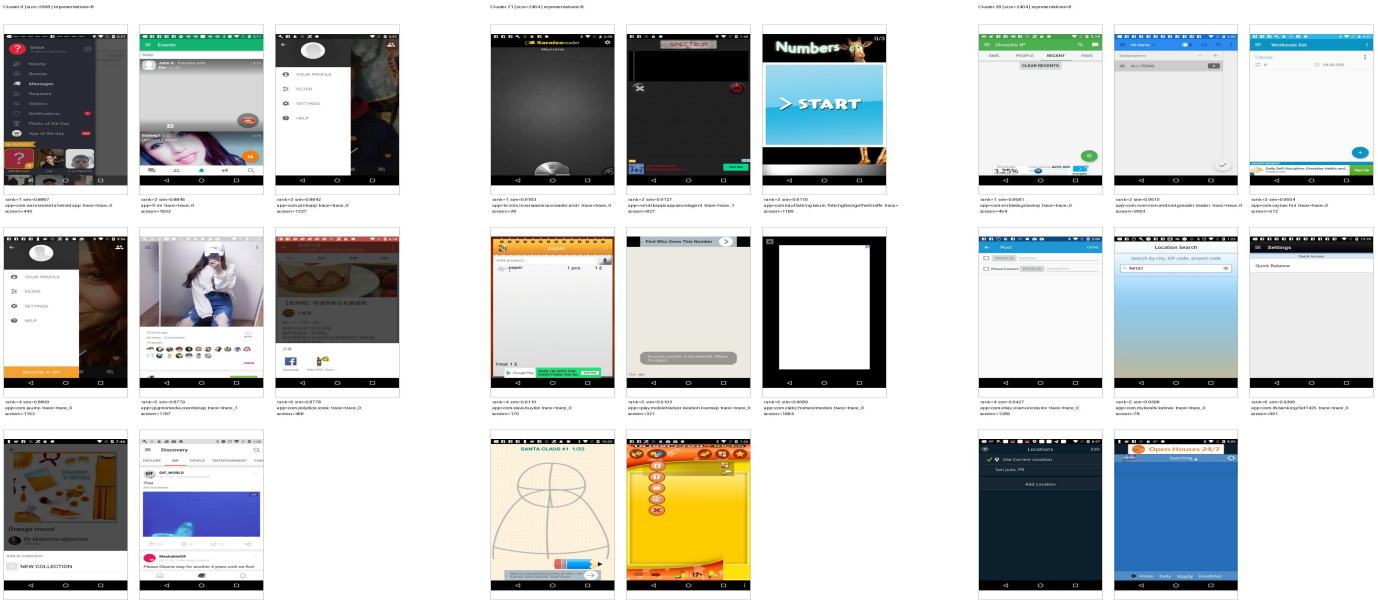


Fig. 9. Qualitative audit of representative learned screen roles at $K = 40$ (Part III).

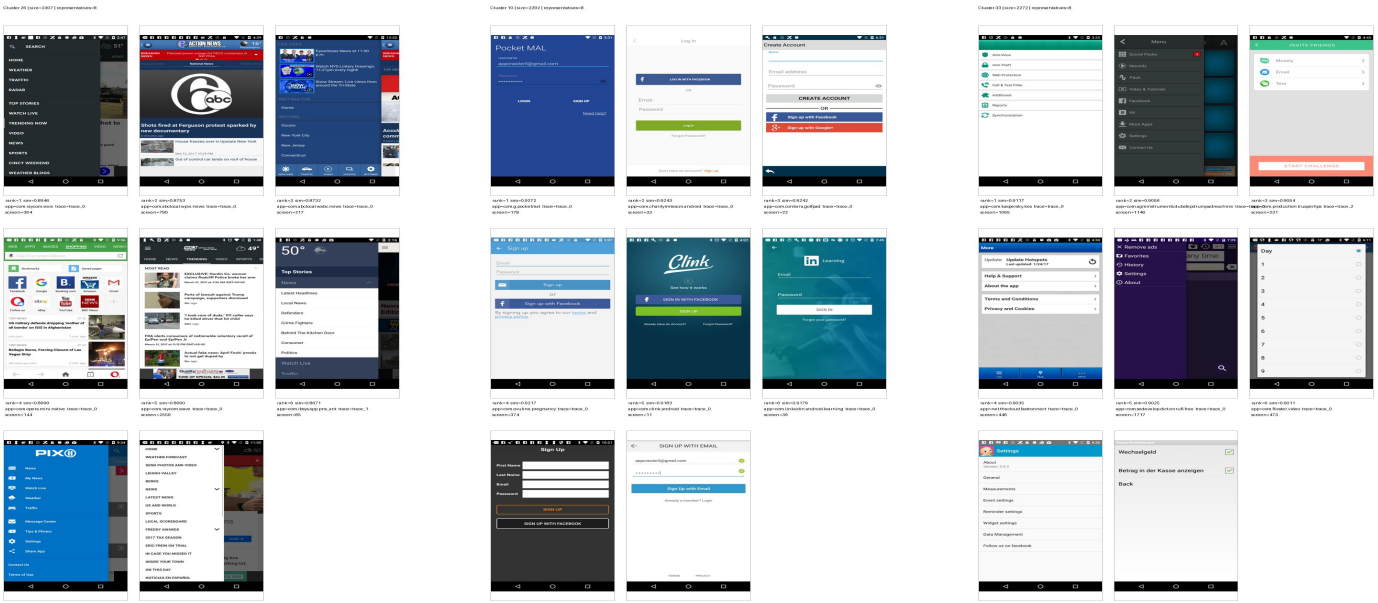


Fig. 10. Qualitative audit of representative learned screen roles at $K = 40$ (Part IV).

Risk Assessment and Risk Management Challenges in Intelligent IoT-Based Smart City Infrastructures

Abdullah Alessa, Yaseen Alduwayl, M M Hafizur Rahman
Department of Computer Networks and Communications-College of Computer Sciences and Information Technology,
King Faisal University, Al-Ahsa, 31982, Saudi Arabia

***Abstract*—Smart Internet of Things (IoT) technologies, which include artificial intelligence (AI) are increasingly being implemented in the infrastructures of smart cities to enhance the efficiency, sustainability, and service delivery of cities. Nonetheless, with the implementation of such intelligent and interconnected systems, there are complex security, privacy, and safety risks that are not easily handled against conventional risk assessment and risk management methods. Current frameworks tend to be unresponsive and centralized whereas smart city infrastructures are dynamic, decentralized, and based on autonomous decision-making elements. This incompatibility poses serious problems for the reliability and robustness of intelligent urban systems. The study contains a systematic literature review of the risk assessment and risk management issues in smart city infrastructure based on intelligent IoT. The review is based on security, privacy, safety, and risks of AI, such as adversarial machine learning, data poisoning, model drift, and failures of autonomous systems. To enhance the level of methodological transparency, the review lists the databases searched, search words, screening process, inclusion, and exclusion criteria, and the resulting list of studies selected. The comparison of the key risk assessment methods, such as the standard-based, qualitative, probabilistic, and AI-based approaches, has been made, and their advantages and weaknesses in the smart city context have been identified. The results indicate that existing frameworks are still in fragments and not always able to deal with the joint effect of the heterogeneity of IoT, AI-based decisions, cyber-physical interdependence, scalability, and governance. The study, based on this summary, offers a more defined taxonomy of risk factors and research directions towards adaptive, AI-conscious and operationally feasible risk management in smart cities.**

Keywords*—*Artificial intelligence; cybersecurity; Internet of Things; risk assessment; risk management; smart cities

ABBREVIATIONS

The following abbreviations are used in this manuscript:

IoT	Internet of Things
AI	Artificial Intelligence

I. INTRODUCTION

The concept of smart cities has become one of the key trends in the current urban development with the fast growth of digital technologies and the necessity to provide efficient, sustainable, and human-centered services. The fundamental aspect of smart city projects is to implement Intelligent Internet of Things (IoT) infrastructures wherein interrelated sensors, actuators, edge devices, and communication networks with artificial intelligence (AI) approaches are used to assist automated monitoring, automated decision-making, and automated control in the urban domain. The systems have become common in transportation management, intelligent energy grids, healthcare services, environmental monitoring, security services, and protection of critical infrastructure [1], [2].

The smart city systems based on intelligent IoT are highly distributed data-driven and adaptive ecosystems compared to traditional urban infrastructures. They keep gathering real-time data using heterogeneous sources and use machine learning and optimization algorithms to facilitate autonomous or semi-autonomous behavior. Although this intelligence has greatly contributed to better operational efficiency and quality of services, there are also complex security, safety, and privacy risks that are totally different to those that exist when the information system is normal [3]. Consequently, this has made proper risk assessment and risk management, an important issue towards long-term reliability and trustworthiness of smart city infrastructures.

A. Background and Motivation

Smart cities are driven by issues that are not limited to the world but are global in nature, including urbanization, population explosion, scarcity of resources, global warming, and the desire to have resilient cities in terms of population services. The amount of governments and municipalities depending on the technologies of IoT and AI to streamline urban processes, cut expenses, and enhance living standards is rising [4], [5]. As an illustrative case, smart traffic management devices dynamically change traffic lights to ease the traffic jams, smart grids regulate energy supply and demand in real-time, and smart surveillance systems assist emergency workers and curb crime with AI-assistance [6], [7], [8].

Nevertheless, the very features that render smart IoT systems appealing also contribute substantially to the scale of the attack on the urban infrastructures. The smart city environments comprise thousands or even millions of heterogeneous devices, with most possessing limited computational power and limited security features. They use wireless and public networks to communicate with each other, and usually, they use cloud and edge computing platforms to process and store data [1], [2]. Such high levels of connectivity and heterogeneity expose people to the risks of cyber threats, including unauthorized access, data breaches, denial-of-service attacks, and system manipulation.

Moreover, AI implementation brings a different category of risks which are not limited to conventional cybersecurity issues. Adversarial attack, data poisoning, model inversion and

concept drift can adversely affect the performance of machine learning models and lead to wrong decisions [9], [10]. Such failures in a smart city setting can cause digital, but also corporeal, damage, in the form of traffic accidents, power outages, incidents involving the safety of the population, and getting crucial services out of order [9], [10].

Conventional risk management models, like the NIST Risk Management Framework or ISO/IEC 27005, were initially intended to be used with fairly stable centralized information systems [11], [12]. These frameworks frequently face difficulties in managing the dynamic, autonomous and decentralized characteristics of intelligent infrastructures based on the IoT. This discrepancy inspires the necessity of a holistic and updated knowledge of the methods of risk assessment and management in the particular context of smart cities.

B. Intelligent IoT-Based Smart City Infrastructures

The concept of intelligent IoT-driven smart city infrastructures may be understood as the macro-scale level of cyber-physical infrastructures based on integrating IoT technologies with AI-powered analytics and decision-making processes. Fig. 1 shows the common notion that these infrastructures are usually comprised of four major layers, namely, the sensing layer, the communication layer, the data processing layer and the application layer [7], [6].

The sensing layer contains physical devices like sensors, actuators, cameras, and an embedded system that is scattered across the city to gather information about traffic flow, air quality, energy consumption, weather conditions and human activities. The communication layer provides the transmission of data based on heterogeneous networking technologies, 5G, LPWAN, Wi-Fi, and wired networks. The data processing layer is based on the edge computing, fog computing, and cloud computing to store and analyze the huge amount of real-time data. Lastly, the application layer provides smart services to city operators, policymakers and citizens via dashboards, automated control system, and decision support systems [1], [2].

The key difference between the intelligent IoT infrastructures and the previous smart systems lies in the wide application of AI methods, including machine learning, deep learning, reinforcement learning, and predictive analytics. The techniques can be used to allow the systems to learn using past and real-time data, adjust to environments undergoing change, and make autonomous decisions with a minimum amount of human involvement [7], [6]. Although such intelligence improves scalability and responsiveness, it also adds complexity to the systems and diminishes transparency, so it is more challenging to predict, measure, and control risks.

C. Problem Statement

Although the intelligent IoT technologies are widely implemented in the smart cities, the analytical interface between the conventional risk assessment methodology and the dynamic nature of AI-driven IoT infrastructures has a great gap. The current risk management techniques are somewhat inertial and theoretical, concentrating on the previously known threats, known vulnerabilities, and fairly fixed system parameters [3].

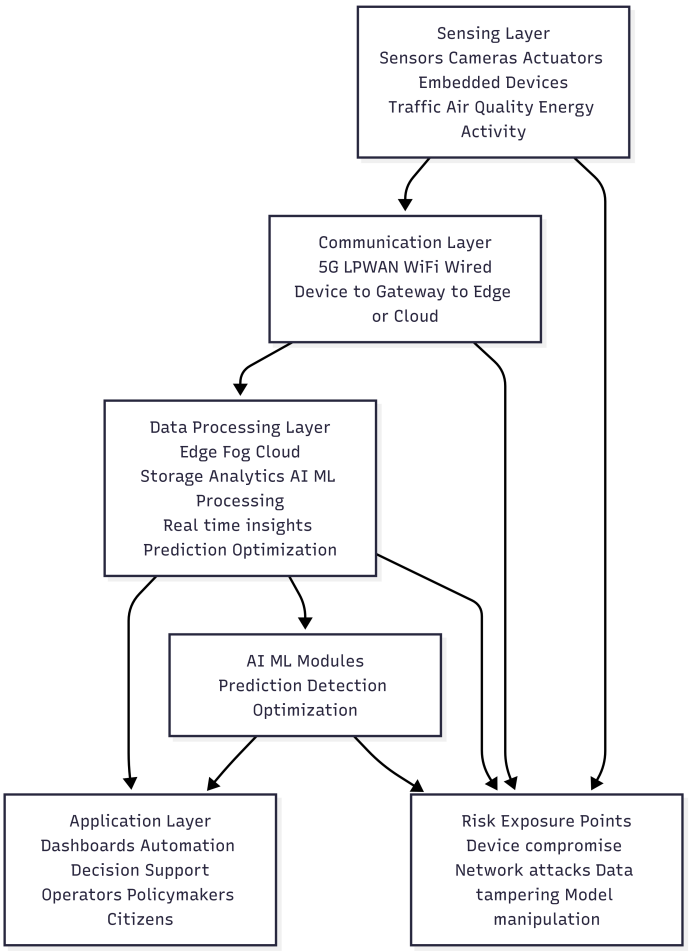


Fig. 1. Layered view of intelligent IoT-based smart city infrastructure, where AI/ML typically operates.

Conversely, smart city systems are also intelligent and they are constantly developing. They are decentralized, use autonomous decision-making algorithms, and are driven by real-time data streams and change of context. The novel and transforming risks brought about by these properties are not properly represented by the traditional models. Specifically, artificial intelligence features like machine learning algorithms present threat on adversarial manipulation, biased training, data poisoning, and unanticipated model behaviour, which may spread within interconnected urban services [9], [10]. As shown in Fig. 2, the reality of intelligent IoT smart city and the traditional risk assessment and management and the gap that addressed between them which reflect on real-world and cause impact to be cyber-physical.

In addition, intelligence IoT failures may have ripple effects on different areas of the city. A subdued traffic management system can affect emergency response, and a smart grid attack can affect the transportation system, healthcare network, and public safety networks all at the same time. This situation where a coherent and consistent knowledge of the way to evaluate and manage such interdependent risks does not exist is a grave threat to the secure and sustainable implementation of smart city infrastructures.

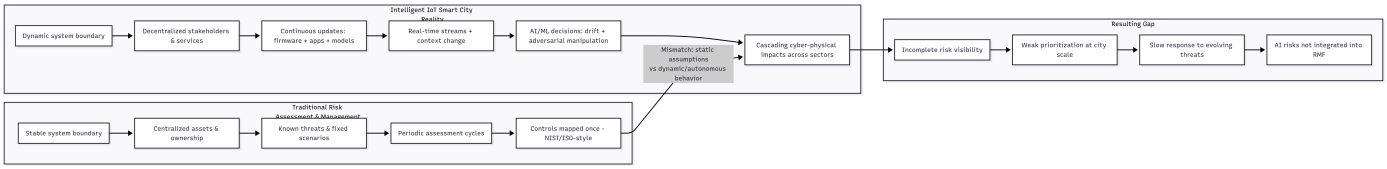


Fig. 2. Mismatch between traditional risk assessment assumptions and the dynamic nature of intelligent IoT-based smart city infrastructures.

D. Research Questions

This comprehensive and rigorous systematic study will address the following research questions to deal with the identified issues:

- RQ1: What are the major security risks and threats specific to the newer AI-based IoT infrastructure within the smart cities?
- RQ2: What are the current risk assessment methodologies and models and what are the technical and practical limitations of the models?
- RQ3: What are the particular difficulties in the process of risk management in real-time, autonomous, and decentralized urban networks?
- RQ4: What are the proposed strategies, frameworks or technologies to reduce intelligent risks in smart city settings?

These research questions will help to organize the study and conduct a full analysis of current methods, problems, and unresolved issues in the area.

E. Contributions of this Review

The review has four key contributions to the literature. First, it introduces a more precise taxonomy of risk factors that impact intelligent IoT-based intelligent city infrastructure, such as technical, AI-specific, cyber-physical, operational, and governance-related risks. Second, it gives a systematic comparison of the significant risk assessment and risk management methods that have been reported in the literature, with a focus on their applicability, strengths, and weaknesses in smart cities. Third, it pinpoints repeating research gaps, particularly in the area of adaptability, scalability, empirical validation, and addressing AI-related threats. Fourth, it outlines practical research opportunities to develop more resilient and context-based risk management approaches towards smart city systems.

F. Paper Organization

The rest of this study will be structured in the following way: Section II outlines the review methodology, which entails the review protocol, databases, search strings, inclusion and exclusion criteria, screening process, quality assessment and data extraction plan. Section III will provide a comparative discussion of the chosen works and summarize the key methodological trends and shortcomings in the literature. The key risk types, risk evaluation methods, risk management issues, and coping strategies are covered in Section IV in the context of intelligent IoT-based smart city infrastructures. Section V provides future research directions. The study ends with a conclusion in Section VI.

II. Review Methodology

The aim of this review was to help identify, analyze, and synthesize recent literature on risk assessment and risk management of intelligent IoT-based smart city infrastructures. The research approach was structured to enhance transparency and reproducibility by having a clear-cut review protocol, database selection plan, screening process, inclusion and exclusion criteria, quality evaluation plan, and thematic synthesis plan. Specifically, the review emphasized the studies that covered cybersecurity, privacy, safety, AI threats, and cyber-physical dependencies of smart cities.

A. Review Protocol

A review protocol was defined before the literature search in order to maintain methodological consistency. The protocol outlined the objectives of the review, research questions, databases to be used, search terms, inclusion and exclusion criteria, screening steps, and quality assessment criteria. The review process was informed with the existing review reporting practices that included transparent reporting of study identification, screening, eligibility assessment and final inclusion. This was done to minimize selection bias and so as to allow a better understanding and reproducibility of the review. Fig. 3 presents the review pipeline used for search, screening, quality assessment, extraction, and synthesis.

B. Search Strategy and Databases

The literature search was done in the following academic databases: IEEE Xplore, Scopus, Web of Science, ScienceDirect, SpringerLink, MDPI, and the ACM Digital Library. The choice of these databases was based on their wide scope of peer-reviewed material on smart cities, IoT, artificial intelligence, cybersecurity and cyber-physical systems.

The search focused on studies published between 2020 and 2026. Only peer-reviewed journal articles and conference papers were considered. The search terms were grouped into four concept clusters: smart cities, IoT systems, artificial intelligence, and risk management. Representative keywords included “Smart City”, “Smart Cities”, “Internet of Things”, “IoT”, “Intelligent IoT”, “Artificial Intelligence”, “Machine Learning”, “Risk Assessment”, “Risk Analysis”, “Risk Management”, “Cybersecurity”, “Privacy”, and “Safety”.

Representative Boolean search strings included the following:

- (“Smart City” OR “Smart Cities”) AND (“Internet of Things” OR “IoT” OR “Intelligent IoT”) AND (“Risk Assessment” OR “Risk Analysis”)

- ("Smart City Infrastructure") AND ("Risk Management") AND ("Artificial Intelligence" OR "Machine Learning")
- ("Smart City") AND ("Cybersecurity" OR "Privacy" OR "Safety") AND ("IoT")
- ("Intelligent IoT") AND ("Smart City") AND ("AI") AND ("Risk Management")
- ("Cyber-Physical Systems") AND ("Smart City") AND ("Risk Assessment")

The search strings were adjusted slightly to match the syntax of each database. This strategy was intended to capture both conventional IoT security risks and emerging AI-specific risks in intelligent smart city environments.

C. Inclusion and Exclusion Criteria

In order to achieve relevance and consistency, strong inclusion and exclusion criteria were used in selecting the studies.

Inclusion criteria:

- Publications of less than five years (2020-2026).
- Conference papers or peer-reviewed journal articles.
- Studies on smart cities, intelligent IoTs, or urban cyber-physical infrastructures.
- Research that deals with risk assessment, risk management, security, privacy, or safety concerns.
- Articles written in English.

Exclusion criteria:

- Studies published before 2020.
- Articles not peer-reviewed or edited or tutorials and opinion papers.
- Research centered on conventional iot that does not feature intelligent/AI-related aspects.
- Articles that are not connected with the context of urban or smart city.
- Redundant papers in databases.

The criteria were established to make sure that the studies that were selected were recent, relevant and directly connected to smart city risk assessment and risk management using intelligent IoT.

D. Study Selection Procedure

The selection of studies was done in several steps. To begin with, all databases were accessed and their records were combined into one dataset and any duplicate records eliminated. Second, the titles and abstracts were filtered through inclusion and exclusion criteria. At this point, studies that were evidently not related to smart cities, intelligent IoT systems, or risk-related issues were filtered out. Third, the entire contents of the rest of the studies were evaluated in terms of topicality and methodological appropriateness. The final review set only used studies that had directly answered the research questions.

E. Search Results and Study Selection Summary

The selection of the study is as follows: In the selected databases, a total of 465 records were initially found. Upon exclusion of 105 duplicates, there were 360 studies left to screen in terms of title and abstract. At this phase, 300 records were eliminated due to the out of scope of the review or failed to meet the inclusion criteria. The rest of the 60 studies were evaluated on full-text. Following the full-text review, 17 articles were filtered out, with 9 being not related to IoT or smart cities, 6 articles lacked a pertinent risk evaluation or risk management lens, and 2 articles were either non-English or old. This led to 43 papers included in the final review (see Table I).

TABLE I. STUDY SELECTION SUMMARY

Records identified	465
Duplicate records removed	105
Records screened	360
Records excluded after title/abstract screening	300
Full-text studies assessed	60
Full-text studies excluded	15
Final studies included in the review	45

F. Quality Assessment of Selected Studies

A structured quality assessment checklist was used in order to assess the methodological quality of the chosen works. The checklist will be based on the existing frameworks of quality assessment of studies and will consist of items like:

- Transparency of research purposes.
- Suitability of methodology.
- Sufficiency of source of data and analysis.
- Applicability to smart city risk management based on intelligent IoT.
- Coherence of findings and conclusions.

All the studies were evaluated in relation to these criteria to make sure that high-quality and relevant research only entered the synthesis. Analyses of studies that had low methodological rigor or unknown contributions were not included in the further analysis.

G. Data Extraction and Synthesis

Relevant information in each of the selected studies was collected by use of a standardized data extraction form. Data taken out comprised the year of publication, the purpose of the research, the area of application, the risk assessment or risk management strategy, essential findings, the challenges identified, and the limitations.

Thematic analysis was a qualitative approach that was used to synthesize the extracted data. The studies were categorised according to the shared themes, including types of threats, risk factors, assessment models, management framework, and AI specific issues. This methodology allowed finding patterns, similarities, and gaps within the literature, which allowed conducting the comparative analysis of the research and discuss the future directions and trends of the studies in a structured way.

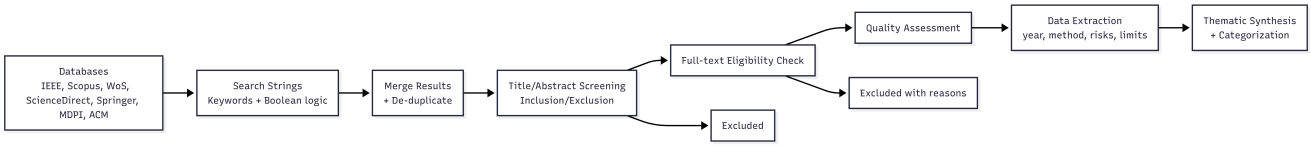


Fig. 3. Review pipeline used for search, screening, quality assessment, extraction, and synthesis.

III. Comparative Analysis of Related Work

This section presents a comparative analysis of the studies selected for the review. Its purpose is to synthesize the literature on risk assessment and risk management in intelligent IoT-based smart city infrastructures, identify common methodological patterns, and highlight unresolved issues.

A. Summary of Reviewed Studies: Objectives and Methods

The study shows that there is an increasing research interest in ensuring smart city infrastructures are secured by risk-sensitive design, as well as risk-sensitive implementation. The majority of the study seeks to determine security threats, to evaluate vulnerabilities, or to suggest frameworks to prevent risks in urban environments with the help of IoT enablers. Nevertheless, they have very different goals and methodology.

Much of the study is dedicated to IoT-based risk assessment, which highlights vulnerabilities at the device level, network security threats, and confidentiality threats to data in the context of smart city implementation [13], [1]. These works often apply conventional security analysis methods, including the threat modeling, attack surface analysis, and qualitative risk matrices, to the situation in urban IoT. Although they are useful in detecting known vulnerabilities, these methods are not always flexible to the dynamic behavior of the system.

The other body of research builds upon classical risk assessment techniques by adding context-sensitive or probabilistic methods, such as Bayesian networks, fuzzy logic, and multi-criteria decision analysis [14]. The purpose of these approaches is to better model uncertainty and interdependencies between smart city elements. Whereas these methods enhance the level of analysis, they are not usually applied at city-wide infrastructures, but applied individually to isolated subsystems, such as smart grids or intelligent transportation systems.

Recent studies are based on applying the methods of artificial intelligence and machine learning to risk assessment and management. These articles address the topic of anomaly detection, predictive risk modelling, and automated threat detection via data-driven methods [15]. Although the above ways demonstrate potential possibilities in managing large-scale and real-time-data, most of the research is still at the concept phase, with minimal implementation in real-world smart city settings.

B. Open Challenges and Research Gaps

A number of open issues and gaps in research can be identified regardless of the variety of approaches to the issue in the study. The first gap is the absence of single risk assessment frameworks that could be used to concurrently consider IoT, AI, and cyber-physical risks related to smart city infrastructures. Fig. 4 illustrate that most of the research is limited to IoT security or AI-based risks, without looking at the potential effect of the two on the interconnected urban systems [16].

The other major problem is that most of the risk assessment models are not dynamic. In traditional and semi-traditional techniques, architecture of the systems and possible attack scenarios are usually fixed and predetermined. Nevertheless, the intelligent smart city systems are constantly changing because of software updates, retraining models, mobility of device, and environmental variations. This time-varying behavior is not well represented in the study models [17], [18].

Scalability is also of great concern. Various suggested frameworks have been shown to be effective in small-scale case studies or simulations, but not in the computational and operational issues of city-wide implementation. The lack of longitudinal research also restricts the knowledge on the transformations of risks over time in smart city intelligent landscapes [1], [2].

C. Key Limitations in Current Research

The comparative analysis indicates several limitations that limit the applicability of the current research in practice. First, in most works, there is rather little attention to AI-specific threats. Although the vulnerabilities of IoT to device compromise and communication attacks are well known, the risks associated with adversarial machine learning, data poisoning, model drift, and biased decision-making are not only briefly mentioned [9], [10] but perhaps not mentioned at all.

Second, the lack of integration of physical safety effects is also a common weakness. Cyber-physical systems are smart city infrastructures, and cyber-attacks may have a direct impact on physical operations and human lives. Nevertheless, most risk assessment studies consider cyber risks individually without clearly modeling the physical or societal impact of such risk [3].

Third, most of the analyzed studies do not present standardized metrics and benchmarks of evaluation, which is why comparing outcomes of various methods can be challenging [19]. This drawback makes the process of selecting appropriate risk management tactics applied to the actual implementation of smart cities more difficult, as well as the formulation of best practices [3].

D. Comparative Analysis of Objectives, Problems, Methods, Findings, and Limitations

To support a structured comparison, the reviewed studies were analyzed across five dimensions: objective, methodology, results, key findings, and limitations [20], [21]. It is noted in

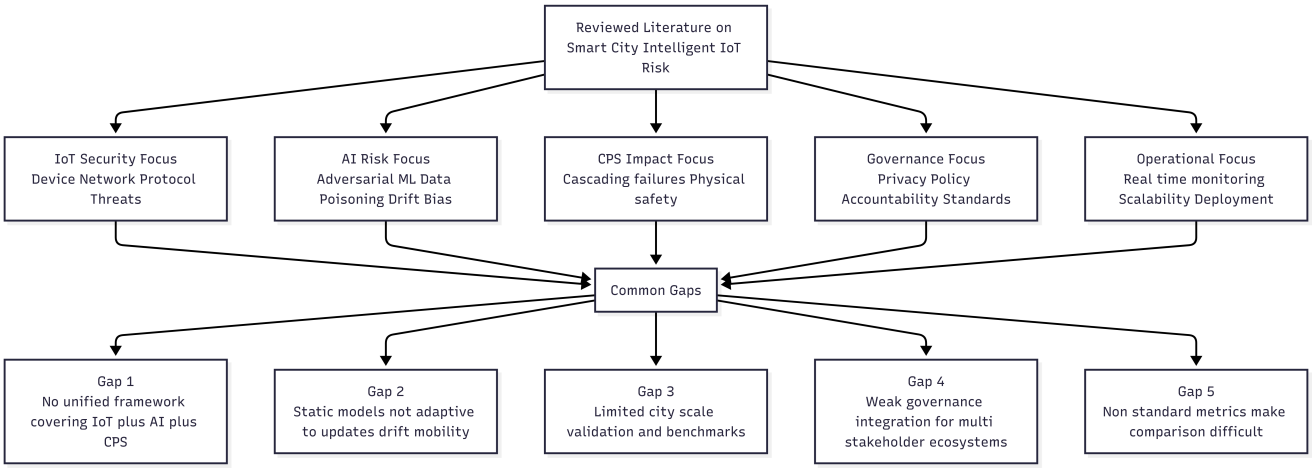


Fig. 4. High-level gap map summarizing limitations across IoT security, AI-specific risks, cyber–physical impacts, governance, and operational scalability in the study.

the analysis that even though the various studies have common objectives e.g. enhance security awareness or minimise the vulnerability of a system, the statement of the problem and the approach used to solve the problem vary significantly among the studies.

Other papers stress on preventive risk management, including secure system design and threat mitigation measures, and others reactive ones, including intrusion detection and incident response [3]. The results of the study tend to establish that intelligent IoT-based smart city systems are confronted with complex risks that cannot be counteracted with unilayer security systems.

E. Synthesized Analysis of Related Work on Risk Assessment and Management in Intelligent IoT-Based Smart Cities

The body of recent research on cybersecurity risk evaluation and management of smart city systems falls under several research strands, such as IoT security, cyber-physical systems (CPS), edge and fog computing, artificial intelligence (AI)-based defenses, governance frameworks, and standardization projects. Initial work on smart city security defines that the heterogeneity, scale and openness of urban IoT ecosystems intrinsically increase the attack surface relative to conventional enterprise systems. According to a foundation of research, smart cities combine sensing, communication, calculation, and control in multi-domain administration, and such risk spread and coordination of attacks are increasingly possible compared to single IoT implementation [22], [4], [23].

Quite a significant study is devoted to the identification of threats and taxonomies based on IoT. Various surveys and examination studies categorize threats to IoT, vulnerabilities, and attack vectors, pointing out vulnerabilities due to constrained devices, insecure communication protocols, weak identity management, and similar [6], [13], [17], [21]. These papers are useful background information but are mostly based on qualitative analysis and taxonomies. Although this sort of approach can be valuable in terms of being aware of a threat, this approach does not provide measurable risk scores regularly updated to support dynamic smart city environments. In addition, they usually consider AI as an external tool and not a part and parcel of the system that is under risk analysis.

The other important research stream deals with risk evaluation structures and methods of smart cities. A number of works can apply classical risk management rules to urban digital infrastructure and focus more on risk identification of assets, the possibility of the threat, and its consequences [24], [20]. Risk management processes and control baselines are further formalized by international standards and guidelines, such as ISO/IEC 27005 and NIST IoT cybersecurity series (NISTIR 8259, NIST SP 800-213, NIST SP 800-53) [12], [11], [1], [16]. Nevertheless, they are very technology-neutral and fixed, and do not provide much information on how risk management can be operationalized in smart city services that are highly adaptive and driven by AI or how to consider the changing threat conditions at a very high rate.

The cyber-physical aspect of smart city risk is considered in CPS-oriented study and demonstrates how cyber attacks propagate into physical disturbances of transport, energy, healthcare, and other, as well as public-safety systems [14], [25]. These publications highlight the harshness of cascading failures and interdependences between digital and physical layers. Nevertheless, the majority of CPS security research is theoretical or computer-based, not being empirically validated, real-time monitoring systems, and quantitative risk prioritization models that can be implemented by the operators of a city.

As edge and fog computing continues to spread, a number of studies explore the effect of decentralized architectures on the smart city threat environment. Mog and edge security surveys determine the new attack vectors of distributed control, data locality, and trust management [26], [27], [28]. Although these papers acknowledge the fact that risk differs at different architectural layers, they do not go further to offer the combined, city-wide risk assessment models that can be used to correlate edge-level events to system-wide impacts.

The use of AI and machine learning in security systems, in particular, intrusion detection and attack detection in IoT networks, is actively studied in recent studies [29], [9], [18]. These methods prove to be better in detecting and being more

flexible than rule-based systems. Nevertheless, they mainly assess performance based on detection-focused measures (i.e., accuracy or false-positive rates) and they do not tend to include adversarial robustness, model drift, or explainability in risk measurement. Introduced by AI as a challenge to the analysis in ENISA reports and threat analysis AI-centered, data poisoning, evasion attacks, and model manipulation have not been dealt with sufficiently in current smart city risk frameworks [7], [30], [10].

Digital twins have been suggested as the way to facilitate the simulation and anticipation of risks at a city scale and provide the opportunity to analyze failures and cyber events in a scenario [19], [8]. Although digital twins provide useful insights on the dynamic processes of complex cities, existing research points out a high level of computational overhead and lack of applicability in real-time. Besides, the majority of digital twins applications focus on operational optimization and do not include constant monitoring and control of cybersecurity risks.

It is also observed that governance, ethics, and sustainability issues are becoming more and more part of smart city cybersecurity. Ethical AI, responsible data use, and sustainable smart city development Studies investigate how technical security controls cannot be effective without open governance and accountability systems [31], [32], [33]. However, such works seldom frame ethical constraints into practice, quantifiable risk management procedures in line with operational cybersecurity interventions.

All in all, the evaluated study shows that there has been substantial development in the domain of threat identification, risk classification, and the suggestion of security systems to be implemented in the IoT systems of Smart Cities, yet, the vital gaps are observed. Current methods tend to be stagnant, inter-technologic, or limited in their coverage of detection performance. There are not many studies that offer combined, quantitative and constantly revised risk management approaches that directly consider AI-driven decision-making, cyber-physical dependencies, and real-time operational constraints. These constraints drive the necessity of a comprehensive risk evaluation and management strategy which is specific to intelligent IoT-based smart city infrastructures namely the ability to incorporate quantifiable risk indicators, AI-conscious threat management, and implementable mechanism to perform ongoing monitoring and mitigation across systems of city-scale systems.

Another study area focuses on security guidance documents, baseline requirements, and regulatory-oriented risk controls related to the implementation of IoT and smart cities. Etis and NIST IoT Core Baseline and other standards and profiles like ETSI EN 303 645 focus on minimum device and system-level capabilities, like secure boot, disclosure of vulnerability, and lifecycle management [34], [8], [35], [36]. Although these documents are necessary to set the basic security hygiene, they are mainly an outline of compliance-based controls as opposed to dynamic risk assessment processes. Consequently, they offer scanty support on the need to prioritize risks, adapt to new threat dynamics, or integrate AI-driven behaviour with current risk management procedures through intelligent smart city infrastructures.

Some works refer to blockchain based and distributed trust systems as a potential facilitator of secure and robust smart city systems. The presence of studies of blockchain integration with IoT and AI reports of enhanced data integrity, decentralization and trust management across heterogeneous stakeholders [37], [16]. Nevertheless, scalability constraints, latency overheads and energy costs that limit real time applicability at city scale are also observed in the surveyed works. Furthermore, blockchain-related solutions tend to center on ensuring the exchange of data as opposed to offering a full-fledged, end-to-end risk assessment platforms that quantify and control cyber and physical risks as well as AI-related risks.

The other new research area is the study of context aware and adaptive security systems of an IoT environment. Context-aware security models seek to adapt protection systems according to the environment, device status or usage patterns, which can enhance responsiveness in smart city contexts [15], [38]. Though they are promising, all these approaches usually run at the control or access management level and have no connection to the higher-level risk assessment processes. They, therefore, add to the local adaptation but fail to offer an uniform perspective of city-wide risk posture or long-term risk trends.

Smart city risk management is indirectly informed through research on resource management and optimization in fog environments and edge environments. Research on the allocation of resources associated with fog and on distributed computing architectures show the impact of workload distribution, latency constraints and resource contention on system performance and resilience [28], [27]. Nevertheless, security and risk considerations do not tend to be a priority or thought of as an assumed limitation; instead, they are often seen as secondary objectives. This lack of correspondence restricts the extrapolability of such works to smart city settings that are intelligent and have resource decisions that can directly affect exposure to cyber and cyber-physical threats.

Lastly, cross-cutting surveys of cyber risk and insurance-focused studies indicate that there is still a multitude of issues with data availability, risk measurement, and intersectoral comparability [3]. Such studies underscore the fact that absence of standardized and high-quality incident and loss data will retard empirical validation of risk models and diminish belief in quantitative risk assessment. It is made worse in smart city settings where a variety of stakeholders, proprietary systems and privacy restrictions further restrict information sharing. Consequently, there has been a high number of suggested smart city risk frameworks that are theoretically tested, but not operationally verified.

Overall, the study confirms that the smart city cybersecurity research is still disjointed on the technical, organizational, and regulatory levels. Although each of the strands is relevant and informative, with some providing background security measures and AI-enhanced detection, others offering governance and ethical aspects, the number of works that combine these viewpoints into a unified, implementable risk assessment and management framework are scarce. This fragmentation highlights the necessity to have an approach whereby AI-aware threat modeling, cyber-physical impact analysis, quantitative risk measures, and ongoing monitoring, are incorporated in

one, operational approach to intelligent IoT-based smart city infrastructures.

Table II shows that the reviewed studies fall into several broad categories: standards-based frameworks, qualitative surveys, probabilistic, CPS-oriented studies, AI-based detection models, adaptive security models, and governance-oriented studies. The analysis of this comparison reveals a definite trend. Standards-based frameworks offer powerful governance framework but are usually inflexible and not responsive to smart city conditions. Taxonomy-based and qualitative studies provide a wide visibility of threats, but seldom assist in dynamic prioritization. Probabilistic and AI-based methods enhance the depth of analysis, but most of them are akin to certain subsystems and are not operationally validated on a city scale. Ethics and governance research brings out issues of accountability and sustainability, yet seldom converts these issues into practicalized risk management processes. Altogether, the literature is still quite fragmented, and there is hardly any research that incorporates under one operational risk management framework the IoT, AI, cyber-physical, and governance dimensions.

IV. Discussion: Challenges and Techniques in Risk Assessment and Management

This section will be the synthesis of the major themes that were identified in the study and are organized in a way that facilitates direct answers to the research objectives and contributions of the current study. The discussion incorporates the technical, operational, and governance perspectives to assess the current state of risk management in smart city infrastructures based on intelligent Internet-of-Things (IoT) technologies.

A. Threat Landscape and Taxonomy of Intelligent Risk Factors

The environment of smart city IoT systems is highly heterogeneous and interconnected where risks come into being on various technical and organizational levels. It is possible to discuss the risk landscape presented in the reviewed literature as a five-layer taxonomy, provided by the study: device layer, network and communication layer, data processing and AI layer, application and service layer, and the governance and organizational layer.

The risks associated with device layers are insecure sensors, embedded devices, firmware vulnerabilities, weak authentication, physical tampering and device compromise. The lack of computational and security power of most IoT devices exacerbates these risks [39], [19].

Examples of network and communication-layer risks are spoofing, running insecure wireless protocols, denial of service, man-in-the-middle and interception of data across heterogeneous communication infrastructures (5G, LPWAN, Wi-Fi, edge-to-cloud).

In automated decision systems, the risks in data processing and AI-layer comprise adversarial machine learning, data poisoning, model inversion, concept drift, biased decision-making, and weak explainability [9], [10], [31].

Risks to application and service-layers involve disruption of smart transportation, energy, surveillance, emergency response and other public services. Due to interconnectedness of these services, the local failure can cause a chain effect on various functions in the city [23], [40].

Governance and organizational risks entail poor accountability, divided risk ownership, poor coordination among the stakeholders, use of nonstandardized metrics, regulatory risk, and inadequate integration of legal, ethical, and operational controls. This stratified taxonomy demonstrates that risk in intelligent smart city infrastructures cannot be confined to single, cybersecurity problems, but rather exists along technical, service, and governance planes (see Fig. 5).

B. Comparative Analysis of Risk Assessment Methodologies

The comparative analysis of the existing study shows a wide range of possible risk assessment models, as shown in Fig. 6, starting with classic qualitative matrices up to the sophisticated data-driven models. The traditional models of qualitative are commonly used due to their simplicity though do not involve the dynamic behavior of decentralized urban ecosystems [3].

More detailed models with the ability to deal with uncertainty and interdependence are available with quantitative and semi-quantitative techniques, including Bayesian networks and fuzzy logic [1], [2], [30]. Nonetheless, these approaches need quality data and in-depth domain expertise, which is often not accessible in advanced smart urban areas. Authors delve more recently into AI-powered anomaly detection and predictive modeling [9], [10]. Although such models hold potential in the real-time monitoring, they have major issues that include transparency, explainability, and susceptibility to adversarial manipulation.

Risk treatment procedures are provided in an organized manner through the use of standard frameworks, e.g. NIST RMF and ISO/IEC 27005 [11], [12]. However, this study finds a large discrepancy: such frameworks assume boundaries of systems and definite ownership of assets, which are not present in the context of decentralized and adaptive smart city conditions [3].

C. Challenges in Real-time, Autonomous, and Decentralized Management

This section summarizes major gaps and constraints that impair efficient risk management in smart city networks. One of the most critical gaps is the lack of a cohesive model that would mitigate the IoT security, AI-specific risks, and cyber-physical implications. Most existing models are not dynamic and do not consider the time varying behavior of the systems that are being continuously updated in software and retraining the models [3].

The two non-technical issues that are dominant are privacy and governance [38]. Smart cities gather sensitive behavioural information and therefore pose a major issue of balancing intelligent service delivery and privacy of data [3]. The absence of specific regulatory frameworks and multiple stakeholders, including the government, the private sector and citizens, also complicate governance because it becomes difficult to

TABLE II. SUMMARY AND COMPARISON OF RELATED WORKS ON RISK ASSESSMENT AND MANAGEMENT IN INTELLIGENT IOT-BASED SMART CITIES

Ref	Objective	Methodology	Results	Key Findings	Limitations
[22]	Analyze security challenges in smart cities	Conceptual analysis of IoT-based city services	Qualitative insights	Heterogeneity increases attack surface	No quantitative risk metrics; IoT-centric
[4]	Study threat modeling for smart cities	Examination study with Bayesian focus	Taxonomy of threat models	Highlights need for probabilistic modeling	No operational validation
[23]	Survey cybersecurity challenges in smart cities	Narrative study	Identified major attack vectors	Emphasizes multi-layer security	Lacks risk quantification
[6]	Study IoT cybersecurity risks	Examination study	Classified IoT risks	Shows fragmented risk handling	Static, non-adaptive
[13]	Taxonomize IoT security risks	Survey and taxonomy	Risk categories identified	Useful baseline taxonomy	No city-scale focus
[17]	Survey IoT cybersecurity techniques	Comprehensive survey	Broad coverage	Highlights attack diversity	No risk prioritization
[21]	Identify critical IoT security categories	Examination study	Security dimensions identified	Structured threat view	Lacks smart city context
[24]	Assess cyber risks in smart city infrastructure	Risk assessment framework	Conceptual risk model	Highlights infrastructure interdependencies	No real-time capability
[20]	Study smart city risk management	Examination study	Identified gaps	Confirms lack of quantitative methods	Study-only
[12]	Standardize information security risk management	Risk management standard	Prescriptive guidance	Widely applicable framework	Static, non-AI-aware
[11]	Define IoT cybersecurity activities	Guideline-based framework	Baseline controls	Strong governance focus	Not smart-city-specific
[1]	Secure IoT devices in government systems	Requirement-based guidance	Security baselines	Improves device hygiene	No risk scoring
[16]	Define security controls for systems	Control catalog	Comprehensive controls	Supports compliance	Heavyweight for real-time use
[14]	Study cyber–physical risks in smart cities	CPS security analysis	Identified cascading risks	Cyber impacts physical systems	Lacks empirical metrics
[25]	Analyze CPS threat landscape	ENISA threat analysis	Threat trends	Confirms CPS risk severity	No city-specific metrics
[26]	Survey edge-based IoT security	Architectural survey	Security design patterns	Edge reduces latency	Fragmented risk handling
[27]	Study edge/cloud IoT computing	Comparative survey	Architecture comparison	Highlights decentralization risks	No risk modeling
[28]	Study fog resource allocation	Examination mapping	Resource strategies	Resource constraints affect resilience	Security treated indirectly
[29]	Detect IoT attacks using ML	ML-based detection	Improved detection accuracy	ML effective for IoT attacks	Detection-only focus
[9]	Study AI-driven IoT security	AI-based threat analysis	Identified AI vulnerabilities	AI introduces new attack surfaces	No risk management integration
[18]	Map AI-based smart city security	Examination mapping study	Technique classification	AI widely used in defense	Ignores AI risk
[7]	Analyze cyber threat landscape	Threat intelligence report	Threat trends identified	Confirms evolving risks	Non-operational
[30]	Identify AI cybersecurity challenges	ENISA analytical report	Risk categories	Highlights adversarial ML	No mitigation framework
[10]	Identify emerging cyber threats	Conference threat analysis	Threat sources identified	AI threats increasing	High-level analysis
[19]	Study city digital twin potential	Examination study	Use cases identified	Supports risk simulation	High computational cost
[8]	Define digital twin requirements	Technical report	Deployment guidelines	Improves system understanding	Limited real-time feasibility
[37]	Study blockchain in IoT & AI	Survey	Security benefits identified	Improves trust & integrity	Scalability issues
[15]	Enable adaptive IoT security	Context-aware models	Improved local adaptation	Dynamic control	No global risk view
[3]	Study cyber risk data availability	Examination study	Data gaps identified	Quantification remains weak	Limits empirical validation
[33]	Link data-driven cities & sustainability	Evidence synthesis	Conceptual insights	Governance is critical	No security metrics
[31]	Explore ethical AI use	Ethical analysis	Normative guidance	Ethics essential for AI systems	Not operationalized
[32]	Study smart vs sustainable cities	Examination study	Sustainability gaps	Security tied to governance	No technical depth

establish risk ownership and accountability [3], [41]. The study concludes that lack of standardised metrics and benchmarks makes it a great challenge to compare the efficacy of various risk management strategies in different city-scale projects.

D. Proposed Strategies and Technologies for Risk Mitigation

In order to overcome the identified threats, the proposed study will use emerging technologies and suggest a pathway to adaptive and resilient infrastructures where Fig. 7 explain the steps of the strategies [1], [2]. Edge and fog computing is referred to as the main enablers in cutting the latency and the exposure of data since information is processed closer

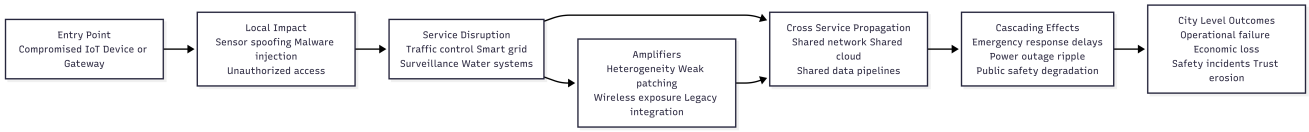


Fig. 5. Example of cascading threat propagation across interconnected smart city services, illustrating how local compromise can trigger city-wide impact.

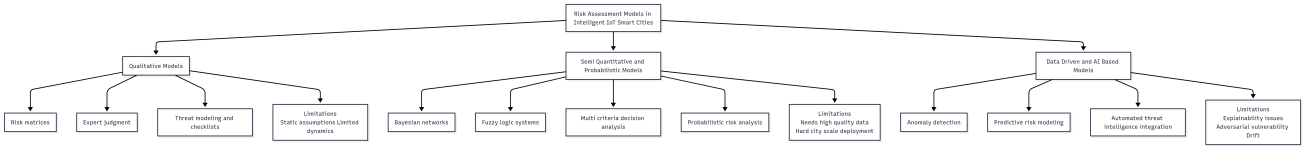


Fig. 6. Taxonomy of risk assessment models reported in the study for intelligent IoT-based smart city systems, including qualitative, probabilistic, and AI-driven approaches.

to the source. Blockchain technologies are suggested to also promote the integrity and auditability of data in distributed environments [1], [2].

The study is more and more in favor of the shift to AI-driven, self-healing security infrastructure that is able to detect anomalies automatically and reconfigure components on the fly [3]. Nevertheless, such solutions further complicate the situation and require stringent validation in order to prevent unexpected outcomes.

Last but not least, this study has added an abstract maturity map, which implies that smart city risk management needs to transform baseline, fixed controls into governance-adopted, AI-conscious, and self-corrective approaches. Risk management cannot be based on one technology, since it should combine technical controls with legal, ethical, and social factors [32].

E. Illustrative Real-World Incidents and Operational Implications

The recorded cases can be used to demonstrate that the risks listed in the literature are not merely theoretical. In the 2018 City of Atlanta ransomware attack, the fundamental operations of the city were disrupted as over a third of the city software applications were offline or partially disabled, including mission-critical services such as police and courts. In 2024, a cyberattack on Transport for London resulted in unauthorised access to a small amount of personal data of some customers, including contact details and potential bank-account information of some users. In 2025, St. Paul, Minnesota, was hit by a coordinated digital attack which the city responded to by shutting down information systems, leading to Wi-Fi disconnection in buildings, library services disrupted, and network resources suspended. Simultaneously, in 2024, U.S. federal government officials advised that disruptive attacks on water and wastewater systems were being conducted by foreign hackers, citing an example in Pennsylvania where a controller at a water facility was incapacitated. These instances demonstrate that the breakdown of digital municipal systems may lead to operational interruption, degradation of the public-services, exposure to privacy, and even safety implications among interconnected urban infrastructures.

V. Future Research Directions

The study also shows that the existing risk assessment and risk management methods are still inadequate to handle the complexity and dynamism of intelligent IoT-based smart city infrastructures. This section provides an overview of some of the most important future research directions that can be used to create more adaptive, resilient, and trustful systems of smart cities. Fig. 8 presents a staged maturity roadmap for evolving risk management in smart cities from baseline controls toward adaptive, AI-aware, and governance-aligned approaches.

A. Adaptive Risk Assessment Models

Creating adaptive risk assessment models that can address continuous modifications in smart city ecosystems is one of the research directions that are of the utmost importance. Future strategies are to be updated dynamically according to the real-time system conditions, environmental factors, and the changing threat intelligence, unlike traditional, inert systems [3].

Continuous monitoring mechanisms and exploiting the streaming data of the IoT devices, edge platforms, and operational logs should be incorporated as adaptive models. The machine learning methods can be used to perceive the new threats and update the risk estimations. Nonetheless, issues of data quality, model drift, and explainability needs to be considered in future studies as well since, in this way, adaptive assessments will be reliable and transparent [1], [2].

B. AI-Driven and Self-Healing Security Frameworks

Self-healing and AI-driven security systems are a good way of controlling risks in autonomous infrastructures of smart cities. These types of frameworks are designed to identify security events, identify root causes, and automatically implement mitigation measures with the least human intervention [3], [35].

Further studies are required in the future to combine AI-based risk management with the resilience principles of systems to allow smart infrastructures of a city to retain satisfactory levels of services even in case of attack or failure. This covers automated isolation of affected parts, dynamic rerouting of network routes, and smart recovery facilities. The

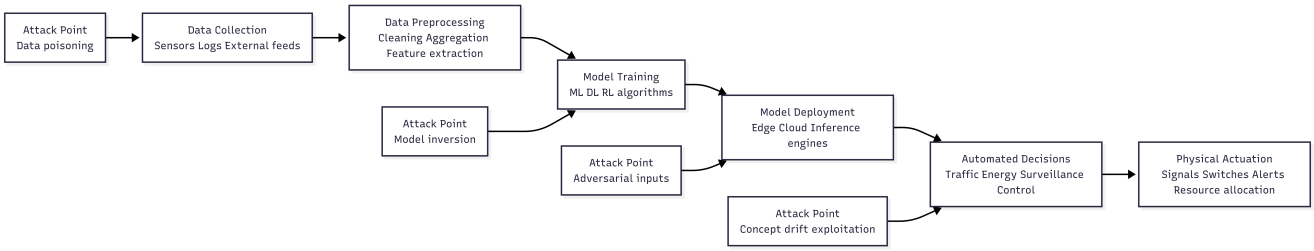


Fig. 7. AI/ML pipeline attack surface in intelligent IoT smart city systems, highlighting adversarial manipulation, data poisoning, model inversion, and drift-related risks.

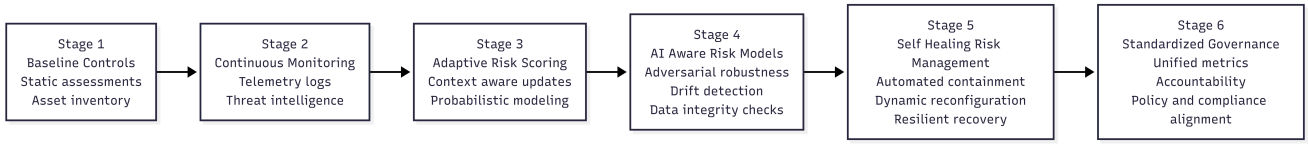


Fig. 8. Proposed maturity roadmap for evolving from baseline risk controls toward adaptive, AI-aware, self-healing, and governance-aligned risk management in smart cities.

research problem of ensuring the strength and safety of these self-healing mechanisms is still open to research [3], [8].

C. Standardization Needs for Smart Cities

Their interoperability and inability to achieve secure smart city deployments are largely due to the absence of standardized risk assessment and management practices. Future studies would help in the establishment of standards specific to domain that deal with the special features of smart IoT-based city infrastructures [3].

The standardization should encompass the technical, organizational, and governance domains, such as risk metrics, threat classification schemes, data governance models, and AI accountability needs. Finding alignment of smart city standards with the existing frameworks including NIST and ISO and extending them to cover AI-related and cyber-physical risks is also a major research priority [11], [12].

D. Practical Deployment Considerations

A gap exists between research and practical implementation to bring about risk management in smart cities. Most of the suggested frameworks and models can be tested by simulation only or through small-scale case studies which restricts their application in practice. The next generation of work would focus on big scale pilot applications and longitudinal designs which would evaluate risk changes over time [3], [42].

Also, the practical implementation should take resource availability, integration of the legacy system, regulatory requirements and coordination of stakeholders. System design should also consider the human factors with the trust of the operator in the risk management tools and the ease of use. These practical considerations will be important issues to address in order to realise the successful implementation of advanced risk management solutions in actual smart city settings [3], [43].

VI. Conclusion

Smart city IoT infrastructures can be extremely effective and efficient, highly automated, and of excellent quality of services, yet they also create complex and changing risks that do not fit the well-established methods of risk evaluation and risk management. This review has explored the recent literature that has been at the crossroads of IoT, artificial intelligence, cybersecurity, and smart city systems. The analysis revealed that current methods are still disjointed. A large amount of research is dedicated to single technical challenges, whereas less is devoted to the interaction of AI-based decision-making, cyber-physical interdependence, complexity of governance, and limitations of operations at the scale of cities.

The review also revealed that the existing frameworks tend to be lacking in flexibility, empirical testing, and standard measures to be used in actual smart city settings. Despite the fact that standards-based models offer a helpful governance framework, and AI-based methodologies enhance detection capabilities, the literature does not report common and operationally feasible frameworks that combine technical, organizational and societal aspects of risk.

The study builds on these results and proposes that future research ought to be directed at adaptive and situation-aware risk assessment model, enhanced treatment of AI-specific threats, enhanced analysis of cyber-physical impact, and more realistic validation in deployed smart city environments. A better taxonomy of risk factors and a more systematic comparison of the current strategies can help build a more robust and reliable smart city infrastructure.

Author Contribution

All authors have equally contributed. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency of Graduate Studies and Scientific

Research, King Faisal University, Saudi Arabia under the [GRANT No. KFU261919].

Data Availability Statement

The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgment

The authors wish to express their gratitude to the Deanship of Scientific Research, Vice Presidency of Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia.

We would like to acknowledge the anonymous reviewers who made great contributions with their brilliant scholarly intuitive comments and sagacious recommendations to improve the quality and clarity of this study.

Conflicts of Interest The authors declare no conflicts of interest.

References

[1] M. Fagan, J. Marron, J. Brady, Kevin G., B. B. Cuthill, K. N. Megas, R. Herold, D. Lemire, and B. Hoehn, "IoT device cybersecurity guidance for the federal government: Establishing IoT device cybersecurity requirements," National Institute of Standards and Technology (NIST), NIST Special Publication 800-213, 2022. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-213

[2] K. N. Megas, M. Fagan, K. Scarfone, and M. Smith, "IoT device cybersecurity requirement catalog," National Institute of Standards and Technology, Tech. Rep. NIST SP 800-213A, 2021. [Online]. Available: https://www.nist.gov/publications/iot-device-cybersecurity-guidance-federal-government-iot-device-cybersecurity

[3] F. Cremer, B. Sheehan, M. Fortmann, A. N. Kia, M. Mullins, F. Murphy, and S. Materne, "Cyber risk and cybersecurity: A systematic review of data availability," *The Geneva Papers on Risk and Insurance - Issues and Practice*, vol. 47, no. 3, pp. 698–736, 2022.

[4] M. Wright, H. Chizari, and T. Viana, "A systematic review of smart city infrastructure threat modelling methodologies: A bayesian focused review," *Sustainability*, vol. 14, no. 16, p. 10368, 2022.

[5] A. Clim, A. Toma, R. D. Zota, and R. Constantinescu, "The need for cybersecurity in industrial revolution and smart cities," *Sensors*, vol. 23, no. 1, p. 120, 2023.

[6] T. S. AlSalem, M. A. Almaiah, and A. Lutfi, "Cybersecurity risk analysis in the IoT: A systematic review," *Electronics*, vol. 12, no. 18, p. 3958, 2023.

[7] M. Barros Lourenco and L. Marinos, "ENISA threat landscape 2019/2020 – the year in review," European Union Agency for Cybersecurity (ENISA), ENISA Threat Landscape Report, Oct. 2020. [Online]. Available: https://doi.org/10.2824/552242

[8] "ETSI en 303 645 v3.1.3 cyber security for consumer internet of things: Baseline requirements," European Telecommunications Standards Institute (ETSI), Tech. Rep., 2024. [Online]. Available: https://www.etsi.org/deliver/etsi_en/303600_303699/303645/03.01.03_60/en_303645v030103p.pdf

[9] W. Zhao and P. Wang, "AI-driven cybersecurity in IoT-based systems," *Sensors*, vol. 25, no. 23, p. 7254, 2025.

[10] J. Simola and T. Leppanen, "Identification of the emerging sources of cybersecurity threats," in *European Conference on Cyber Warfare and Security*, vol. 24, no. 1, 2025, pp. 794–802.

[11] M. Fagan *et al.*, "Foundational cybersecurity activities for IoT device manufacturers," National Institute of Standards and Technology, Tech. Rep. NISTIR 8259, 2020. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8259.pdf

[12] *ISO/IEC 27005:2022 Information Security Risk Management*, International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), 2022. [Online]. Available: https://blog.ansi.org/ansi/iso-iec-27005-2022-security-risk-management/

[13] M. Waqdan, H. Louafi, and M. Mouhoub, "Security risk assessment in IoT environments: A taxonomy and survey," *Computers & Security*, vol. 154, p. 104456, 2025.

[14] M. Houichi, F. Jaidi, and A. Bouhoula, "Cyber security within smart cities: A comprehensive study and a novel intrusion detection-based approach," *Computers, Materials & Continua*, vol. 81, no. 1, pp. 393–441, 2024.

[15] M. Fagan *et al.*, "IoT device cybersecurity capability core baseline," National Institute of Standards and Technology, Tech. Rep. NISTIR 8259A, 2020. [Online]. Available: https://csrc.nist.gov/pubs/ir/8259/a/final

[16] J. T. Force, "Security and privacy controls for information systems and organizations, revision 5," National Institute of Standards and Technology (NIST), NIST Special Publication 800-53, Revision 5, 2020. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-53r5

[17] E. Dritsas and M. Trigka, "A survey on cybersecurity in IoT," *Future Internet*, vol. 17, p. 30, 2025.

[18] R. Alraddadi, M. Alshayeb, S. Mahmood, and M. Niazi, "Enhancing cybersecurity in smart and cognitive cities: A systematic mapping of AI-based techniques," *Array*, vol. 28, p. 100606, 2025.

[19] E. Shahat, C. Hyun, and C. Yeom, "City digital twin potentials: A review and research agenda," *Sustainability*, vol. 13, p. 3386, 2021.

[20] J. Alshehri, A. Alhamed, and M. M. H. Rahman, "A systematic literature review on cybersecurity risk management in smart cities," in *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Osaka, Japan, 2024, pp. 407–412.

[21] H. Sebestyen, D. Popescu, and R. Zmaranda, "A literature review on security in the internet of things: Identifying and analysing critical categories," *Computers*, vol. 14, p. 61, 2025.

[22] C. K. Toh, "Security for smart cities," *IET Smart Cities*, vol. 2, no. 2, 2020.

[23] J. S. Oliha and P. W. Biu, "Securing the smart city: A review of cybersecurity challenges and strategies," *Open Access Research Journal of Multidisciplinary Studies*, vol. 7, no. 1, pp. 094–101, 2024.

[24] M. Kalinin, V. Krundyshev, and P. Zegzhda, "Cybersecurity risk assessment in smart city infrastructures," *Machines*, vol. 9, no. 4, p. 78, 2021.

[25] M. Goncalves *et al.*, "ENISA threat landscape 2021," ENISA, Tech. Rep., 2021. [Online]. Available: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2021

[26] K. Sha, T. A. Yang, W. Wei, and S. Davari, "A survey of edge computing based designs for IoT security," *Digital Communications and Networks*, vol. 6, no. 2, 2020.

[27] F. Andriulo, M. Fiore, M. Mongiello, E. Traversa, and V. Zizzo, "Edge computing and cloud computing for internet of things: A review," *Informatics*, vol. 11, p. 71, 2024.

[28] I. B. Lahmar and K. Boukadi, "Resource allocation in fog computing: A systematic mapping study," in *2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC)*, Paris, France, 2020, pp. 86–93.

[29] M. F. Hyder, W. Nazir, M. U. Farooq, M. Anwer *et al.*, "Attack detection in IoT using machine learning," *Engineering, Technology and Applied Science Research*, vol. 11, no. 3, pp. 7273–7278, 2021.

[30] European Union Agency for Cybersecurity (ENISA), "Artificial intelligence cybersecurity challenges," ENISA, Tech. Rep., 2020. [Online]. Available: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

[31] M. Taddeo and L. Floridi, "How AI can be a force for good," *Science*, vol. 361, no. 6404, pp. 751–752, 2020.

[32] T. Yigitcanlar, L. Kamruzzaman, M. Foth, and J. Marques, "Can cities become smart without being sustainable? a systematic review of the literature," *Sustainable Cities and Society*, vol. 45, 2025.

[33] S. E. Bibri, "Data-driven smart sustainable cities of the future: An evidence synthesis approach to a comprehensive state-of-the-art literature review," *Sustainable Futures*, vol. 3, no. 1, p. 100047, 2021.

[34] "ETSI en 303 645 v2.1.1 cyber security for consumer internet of things: Baseline requirements," European Telecommunications Standards Institute (ETSI), Tech. Rep., 2020. [Online]. Available: https://www.etsi.org/standards#303645

[35] M. Gencer *et al.*, "Profile of the IoT core baseline for consumer IoT products," National Institute of Standards and Technology, Tech. Rep. NISTIR 8425 (Initial Public Draft), 2022. [Online]. Available: https://csrc.nist.gov/external/nvlpubs.nist.gov/nistpubs/ir/2022/NIST.IR.8425.ipd.pdf

[36] M. Fagan *et al.*, "NISTIR 8259 series: Cybersecurity for IoT program," National Institute of Standards and Technology, Tech. Rep., 2021. [Online]. Available: https://www.nist.gov/itl/applied-cybersecurity/nist-cybersecurity-iot-program/nistir-8259-series

[37] H. F. Atlam, M. A. Azad, A. G. Alzahrani, and G. Wills, "A review of blockchain in internet of things and AI," *Big Data Cogn. Comput.*, vol. 4, p. 28, 2020.

[38] T. Xue, Y. Zhang, Y. Wang, W. Wang, S. Li, and H. Zhang, "Edge computing for IoT: Novel insights from a comparative analysis of access control models," *Computer Networks*, vol. 270, p. 111468, 2025.

[39] G. M. Abreu, A. Pan, A. E. Post, N. Malkin, and K. T. Frick, "How do cyber-risks vary across smart city technologies?" *Journal of Urban Technology*, pp. 71–91, 2025.

[40] B. Sereda and J. Jaskolka, "An evaluation of IoT security guidance documents: A shared responsibility perspective," *Procedia Computer Science*, vol. 201, pp. 281–288, 2022.

[41] M. Alowaidi, S. K. Sharma, A. AlEnizi, and S. Bhardwaj, "Integrating artificial intelligence in cyber security for cyber-physical systems," *Electronic Research Archive*, vol. 31, no. 4, pp. 1876–1896, 2023.

[42] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *Systematic Reviews*, vol. 10, no. 1, p. 89, 2021. [Online]. Available: https://link.springer.com/article/10.1186/s13643-021-01626-4

[43] N. Rieke, J. Hancox, W. Li, F. Milletari *et al.*, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, no. 1, 2020.

A Cryptographic Framework Using AES-ECC with Threshold Key Management for Cloud Storage Systems

Abdulsalam Ibrahim Almirdasi, Mohamed Tahar Ben Othman
Department of Computer Science-College of Computer, Qassim University, Saudi Arabia

***Abstract*—Cloud storage systems have become an essential platform for storing and managing large volumes of data, but their security depends not only on confidentiality but also on integrity, controlled key management, and resistance to active attacks. Many existing protection approaches emphasize encryption of data while giving less attention to context–aware verification and controlled object recovery in untrusted cloud settings. This study proposes a novel hybrid cryptographic model for cloud storage systems, Object-Centric Threshold–Sealed Encryption with Two Keys (OCTET). The model integrates AES chunk-based encryption to protect data confidentiality, ECC to secure key exchange, threshold–based secrets to reconstruct, HKDF–derived per chunk, and Merkle root to enforce a verification–before–decryption policy. The implementation of the proposed model is in an emulated cloud storage system and examined on a dataset of large objects, then compared against baseline schemes, including symmetric and hybrid encryption models, under the same experimental environment. The main outcome demonstrates that the proposed model achieves practical performance, minor overhead, and superior resistance to the attack models. In general, this study demonstrates that the proposed model offers trade–offs between security and efficiency and a robust integrity technique for large objects in cloud storage systems.**

***Keywords*—Cloud storage system; hybrid encryption; AES; ECC; Object-Centric; threshold key management; integrity verification; Merkle root verification**

I. Introduction

Cloud computing systems are essential for providing trustworthy computation and storage resources, allowing organizations to easily manage data without using local servers [1]. This way of data storage brings several challenges related to potential data access and misuse [2]. While cloud solutions can be attractive for their flexibility, cost savings, and easier management, they also come with certain drawbacks, such as challenges related to multi–tenancy and having less direct control over the underlying hardware [3].

In addition, complicated service–based agreements and an external attack surface adds more complicated security. Any cloud data protection plan must, therefore, carefully balance these benefits against the risks involved. Although cloud–based systems provide advantages like flexibility, effectiveness, and enhanced control, these systems present challenges, such as multi–tenancy issues and limited hardware control [3]. Furthermore, complex agreements of service and a broad attack surface complicate security [2]. Any plan of data protection in cloud must balance these advantages and challenges.

Cryptographic schemes serve as a vital defense technique to secure data on the cloud and they use three different methods to provide security, which are symmetric, asymmetric, and hybrid key encryption. Symmetric methods utilize one key to encrypt/decrypt data, applying different schemes such as AES, Blowfish, and 3DES [1]. They are considered a viable option for handling large data because of their efficiency in rapid encryption. In the other hand, asymmetric encryption uses two different keys, where the public secret encrypts data and the private secret decrypts data, such as RSA and ECC, making it a proper option for key exchange due to their security robustness [4]. The hybrid models combine both methods, with a symmetric part for encrypting the large data and an asymmetric part to secure key exchange [5]. In the context of data encryption algorithms for cloud storage systems, it a viable choice to select hybrid encryption methods to provide the framework with fast data encryption by using a symmetric technique and secure key exchange between channels by using an asymmetric technique.

Furthermore, key management is an important concept in cryptography. If the keys are not effectively exchanged, stored, or generated, it means that the implementation of the cryptographic scheme is ineffective and cannot function successfully [1]. In cloud storage systems, keys should be generated with efficient randomness, linked with certain characteristics, and tracked via whole lifecycle. However, threshold–or distributed–based environments, key servers, and hardware security modules (HSM) each has distinguish advantages and disadvantages in terms of security, complexity, and availability [4]. As a growing number of organizations adopt cloud–based services, they need to carefully assess both their advantages and disadvantages to ensure that strong security measures are established. Understanding the main concepts and ideas of each strategy may significantly impact how efficient data security is.

For protecting cloud storage, there are some security concepts, which are confidentiality, integrity, and availability (CIA). Confidentiality is basically when authorized clients are able to read and access the data in cloud storage [4]. Cloud systems use cryptographic techniques and keeps the keys from being released to ensure confidentiality. Integrity means that the cloud storage model checks that the data objects have not been changed or deleted while they are stored [6]. The cloud storage model does integrity verification. Users frequently apply techniques such as authenticated encryption, hashing, and Merkle tree commitments to achieve integrity. The availability mainly is when real clients can retrieve and recover their data, whether there is an incident or partial leak

[3].

The main problem of this study is that many modern security techniques emphasize confidentiality instead of focusing on integrity and contextual relevance. It is difficult to manage keys and efficiently store data in cloud storage systems, where attacks commonly occur against users. CSPs must protect sensitive data and maintain the users' accessibility.

The aim of this study is to propose a new hybrid encryption model for cloud storage systems, which is called Object–Centric Threshold–Sealed Encryption with Two Keys (OCTET). The model encrypts data using AES and secures key exchange using ECC to balance between efficiency and security. The framework also has security features like chunk authentication and a secret sharing threshold. This study presents an evaluation and comparison of proposed model against baseline schemes in the context of security and efficiency. The following objectives are defined to achieve the main goal of this study:

- Develop OCTET, a hybrid encryption model that combines AES and ECC to enhance security and integrity through an object–centered approach.
- Adopt advanced security measures, including authenticated chunking, HKDF key derivation, and threshold secret sharing, to safeguard confidentiality, uphold integrity, and regulate access at both object and chunk levels.
- Simulate a reproducible OCTET prototype for cloud storage to assess its resilience against tampering and context–binding attacks, while comparing its efficiency and security with baseline schemes.

The main contribution of this study is to introduce OCTET, a cloud storage security framework that employs chunk encryption to maintain data integrity and user access control. By assigning specific secret keys to tenants and verifying data integrity before decrypting it, OCTET makes unauthorized access more complex. It combines several features, such as AES, ECC, and Shamir secret sharing, to properly manage systems with multiple users. Furthermore, it evaluates the efficiency and security trade–offs.

II. Literature Review

In this section, a review investigates the security of data in cloud computing systems, covering threats like violation of integrity and leakage of data. This review covers the symmetric and hybrid models, as well as the efficiency and security of these cryptographic models. Furthermore, it highlights existing frameworks and their limitations in the context of confidentiality, integrity, and key management flexibility.

A. Cloud Storage Security Challenges

Challenges of security in cloud computing arise due to the lack of control over infrastructure, applications, and data by users, which are handled by cloud service providers (CSP) via different locations. This way may cause some issues, such as loss of data, access for unauthorized users, delays, corruption, and potential risks for sensitive data. In [7], a review addresses the challenge of data leakage, which occurs due to vulnerabilities like cross attacks and admins that maliciously threaten confidentiality and integrity, highlighting the need for security in monitoring reliable computing and teamwork, while also investigating the dependence on new threats and untrusted systems. Also, in [8], another review highlights the significance of CIA as risks arise with Internet of Things (IoT) and cloud use, emphasizing techniques including access control and encryption. Furthermore, in [9], a comprehensive review shows that clients are losing direct control in cloud data centers, which leads to vulnerabilities, such as theft of data and disconnections of service. This demonstrates that more verifiable methods are needed to improve security robustness and efficient data protection.

These studies highlight persistent security issues in cloud storage, including data leaks, insider threats, or tenant attacks that influence CIA. Security measures such as encryption, access control, DLP, blockchain, and redundancy complicate business operations and increase costs. There is no single solution ensuring complete safety, and the proposed models reveal a research gap for integrated cloud–based security models that effectively combine strong encryption, key management, verifiable integrity, and performance efficiency.

These studies mainly define the landscape of threat at an overall level, like access authorization, data leakage, and control loss without providing cryptography designs that combine integrity and confidentiality with controlled recovery for stored data. Therefore, the proposed model (OCTET) is designed to cover these challenges via context-aware verification, threshold key management, and protection of objects without dealing with these security demands separately.

B. Existing Encryption Models for Cloud Storage

Recent trends in improving the security of cloud storage have resulted in different encryption schemes to preserve privacy while enabling search, access control, and other performance improvements. Several studies in the area of encryption models have significantly improved the security and accessibility of cloud data. In [10], SES–CSE enhances data confidentiality and availability by enabling efficient keyword searches in large encrypted cloud datasets while protecting plaintext information and demonstrating resilience against ciphertext and chosen–keyword attacks. Additionally, in [11], ECSES and ECRM make data more secure and available in cloud systems by using strong symmetric encryption and backup methods to lower the chances of problems from node failures or security breaches. Furthermore, in [12], an improved CP–ABE model makes it easier to control who can access encrypted cloud data in IoT environments, ensuring that only authorized users can get to the data and reducing security risks related to hashing functions and signature checks.

In summary, the studies demonstrate that encryption algorithms have been improved over time under cloud storage systems. Currently, encryption frameworks started using extra complicated tools such as searchable methods, redundancy–based check techniques, and attribute–based access control. So, the studies indicated the significance of cryptographic frameworks in terms of security of systems, protection of data, and access control policies.

Significant functions like confidentiality, backup support, searchable techniques, and access control are addressed as improved methods for cryptographic models. However, these functions are discussed as separate solutions without integrating chunking authentication with context-binding and data integrity. Furthermore, the proposed model (OCTET) is built to cover these limitations by using techniques such as AES for chunking and HKDF-derived keys with verification by Merkle root.

C. Hybrid Models and Key Management Mechanism

When combined with key management operations, hybrid encryption algorithms can efficiently alleviate the shortcomings of single cryptographic schemes in cloud computing. Such techniques generally utilize symmetric encryption to protect data while using asymmetric or distributed approaches for key management. Hybrid and key encryption models play a crucial role in enhancing data security in cloud environments. In [13], a blockchain–based key security management framework is proposed that utilizes an oblivious pseudo–random function to safeguard encryption keys in cloud storage, enhancing recovery from failures and ensuring immutable transaction records while demonstrating resilience against brute–force and collusion attacks with low computational costs for key generation. Furthermore, in [14], a hybrid model enhances cloud data protection by combining symmetric and asymmetric encryption algorithms with a private blockchain, utilizing AES for large files and RSA for securing the AES key, thereby improving key exchange and access monitoring. Finally, in [15], the SymECCipher framework combines ECC with AES to efficiently secure healthcare data in the cloud, enhancing key management and throughput protection tailored to the specific needs of cloud healthcare applications.

Finally, these investigations demonstrate the principles of how the proposed model functions. The main focus is that it is essential to supplement fast symmetric with asymmetric or decentralized key management. They also emphasize the advantages of threshold and secret–sharing schemes for secure key backup. The studies also show how blockchain technology or other equivalent techniques can be used to create a decentralized and tamper–proof way to monitor key management and access events in cloud storage systems.

These studies ensure the significance of integrating symmetric and asymmetric methods but still concern secure key exchange, key monitoring, and recovery instead of the combination of protection of chunks, key release, and verification-before-decryption policy. Therefore, key management techniques are improved, while data integrity and recovery control are not imposed simultaneously. On the other hand, the proposed model (OCTET) addresses this issue by using both AES and ECC with chunk-based derivation and threshold methods.

D. Integrity Mechanisms in Cloud Storage

Cloud storage involves transferring data from local devices to third–party servers, necessitating user trust in these providers to maintain data safety and accessibility. Concerns about data manipulation and loss are addressed through cryptographic integrity checks. Recent studies have explored innovative integrity mechanisms designed to bolster data verification and security across various domains. In [16], a study introduces a hierarchical Merkle hash tree for integrity auditing that improves data verification while reducing audit operations and expenses. Also, in [17], a model for checking the integrity of electronic health records (EHRs) in the cloud allows for fast checks without downloading all the data, using cryptographic tags to verify blocks and letting computations be done elsewhere to lessen the load on servers. Lastly, in [18], different storage–layer integrity techniques and the security problems they pose are grouped together, where the focus is on risks like insider threats and ways to improve efficiency.

Recent trends highlight a transition from basic checksum verifications to advanced cryptographic auditing methods that enhance efficiency and security. Hierarchical Merkle hash trees reduce audit costs and effectively detect tampering. Several studies establish methods that balance between vulnerabilities and costs of computations. In general, observations support cryptographic frameworks that provide greater security for large data stored on the cloud while maintaining performance.

These studies demonstrate that Merkle root and verification checks can detect tampered data and decrease the cost of verification. However, these techniques are mainly presented as mechanisms for checks and audits that come on the storage layer, where they are unable to show how integrity checks may be tied with the derivation of keys and control of decryption inside a cryptographic model. Therefore, the proposed model (OCTET) solves this gap by designing a step that ensures integrity before the decryption process using Merkle root verification and checking header authentication to prevent tampered data from the recovery process.

E. Limitations of Current Encryption Models

Although tools of cryptography in cloud environments are notably improved, they still lack solid approaches that encompass performance, supported cross–platform scalability in key management, and integrity assurance. Some of the difficulties addressed by security and encryption techniques are system robustness and user access control balance with performance. In [19], a blockchain-based model is proposed for data security in cloud environments using the combination of AES and RSA for encryption. The model is built using decentralized key management contracts and agreements to enhance confidentiality. Furthermore, in [20], the authors address security techniques using attribute–based cryptography, homomorphic encryption, and authentication of multi-factors. In addition, in [21], a hybrid cryptographic model is proposed using AES with OTP and RSA for cloud computing systems, which establishes dynamic key management with strict access control.

Overall, with the advancement of cryptographic models using blockchain methods, there are still challenges like processing delays and compatibility issues. Also, multi–layer cryptography systems faces challenges such as the increment of costs and complexity in key management. For hybrid models, they improve access control, yet their encryption process caused overhead data, which influenced the system performance.

Generally, these studies show that cloud encryption frameworks improve several aspects like access control, key manage-

ment, and confidentiality, while still lacking a balance between overhead, compatibility, integrated integrity verification, and scalability. Many encryption algorithms are strong in function isolation but still offer limited support for trade-offs in efficiency and security. On the other hand, the proposed model (OCTET) brings a mixture of techniques and behaviors as one cryptographic hybrid model for cloud storage systems.

III. Methodology

To propose a novel cryptographic model, OCTET, for cloud storage systems, the methodology of this research is a mixed–method approach. Qualitative method reviews current studies in cloud security and data encryption schemes to find out design demands. Quantitative method conducts the development and evaluation of OCTET, as well as the experiment setup with a real–world dataset. In terms of efficiency and security, a comparative analysis is also presented to compare the proposed model against baseline schemes.

A. Design of the Proposed Model

The Object–Centric Threshold–Sealed Encryption with Two Secrets (OCTET) model is a hybrid cryptographic framework designed to secure large data in cloud environments. It achieves its goal by binding data to its specific cloud location while employing dual secrets, threshold key release, and verifiable integrity. Firstly, OCTET splits data objects into several chunks of fixed size, then labels each chunk to be identified easily, which makes these labels linked to their corresponding encrypted chunks and linked to valid cloud identities, thus maintaining data security and integrity at its specific location.

OCTET generates two independent secret keys: the root object secret (ROS) and a second secret known as "Delta" (δ), which is deterministically derived from a key management system (KMS). ROS is stored using Shamir–style threshold sharing, allowing it to be reset through elliptic curve cryptography (ECC) after receiving a predetermined number of approvals (k–of–n). Both ROS and δ are processed through a hash–based message authentication key derivation function (HKDF) to produce advanced encryption standard (AES) parameters for each data chunk, preventing key reuse. AES encrypts per–chunk, resulting in ciphertext and a tag. The proposed model builds a Merkle tree from chunks of ciphertext hashes, creating one MerkleRoot that ensures the overall integrity of the object. When clients retrieve encrypted data (ciphertext), the framework reconstructs ROS and calculates δ, validating the Merkle root and header message authentication code (MAC) before it performs decryption, where the header is created containing Merkle root and other parameters, as well as deriving the two secret keys (ROS and δ) by MAC. Upon successful verification, keys of decryption are generated via HKDF, then chunks are decrypted, and objects are recovered into original plaintext. Fig. 1 illustrates the whole structure of proposed model (OCTET), showing the encryption and decryption workflows organized around chunking, key security, and integrity check, which clearly defines that the model is not retrieving the data, while it initially reconstructs the needed key material and verifies the integrity information.

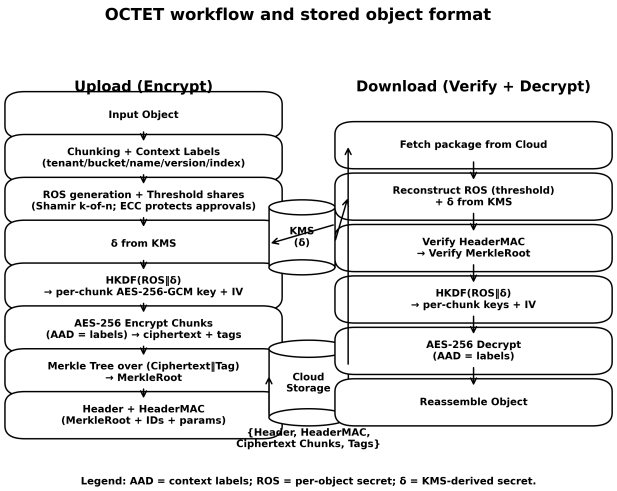


Fig. 1. Encryption and decryption in proposed model (OCTET).

B. Selected Encryption Algorithms

This study presents a comparative analysis of algorithms to assess OCTET for cloud data storage, focusing on key requirements such as data privacy, integrity verification, and key management. The study involves tested algorithms, including hybrid models that offer AEAD and symmetric schemes that utilize an encrypt–then–MAC scheme with HMAC–SHA256 for integrity verification. Under controlled scenarios, the evaluation is presented with consistency using the same settings and workloads for all the selected encryption algorithms to ensure fairness. The algorithms' configurations are listed below to recognize the procedures and roles for the comparison:

1) Proposed Model (OCTET): It employs a 60 MB chunk size for data encryption, utilizing AES–256–GCM for encryption and ECC P–256 for key encryption, which is linked to the object's storage context. Implements a 2–of–3 threshold to reconstruct the ROS per object with the secondary secret key δ, derived during the decryption process.

2) Blowfish: It is configured with a key size of 256 bits and an initialization vector (IV) with a size of 64 bits per chunk, which ensures the consistency for the scenarios measuring the efficiency of the encryption/decryption process.

3) Triple DES (3DES): It applies 3DES–EDE procedures using a setup of two secrets, providing the mode of CBC with 112 bits and PKCS#7 with 8 bytes.

4) Hybrid AES-RSA (AES-RSA): It utilizes AES–256–GCM with a tag size of 128 bits for encrypting data, while RSA–2048 is used to secure the keys of AES with RSAES–OAEP and SHA–256 to encrypt keys.

5) Hybrid AES-ECC (AES-ECC): It also utilizes AES–256–GCM with a tag size of 128 bits for encrypting data, while ECC (P–256) in the secp254r1 curve is used to secure key exchange with ECDH mode integrated with HKDF–SHA256 to derive keys.

The selected schemes offer a standard measurement for evaluation purposes of the proposed model (OCTET). The symmetric algorithms, including Blowfish and 3DES, are used

to present traditional cryptography behavior. The asymmetric algorithms are also selected to present the cryptography behavior. AES–RSA is chosen as a standard hybrid metric that shows the common process, where AES is utilized for the encryption of large data and RSA–OAEP is utilized for key-wrapping security. Additionally, AES–ECC is selected as the nearest hybrid model to OCTET due to both models relying on AES and ECC, which makes it a basic representation for the proposed model. However, AES–ECC is unlike OCTET because of the additional techniques in OCTET, like context-binding, the verification-before-decryption policy, and the threshold method.

C. Data Collection

In this study, the dataset is used for evaluation of OCTET under scenarios for cloud storage of bulk data instead of presenting all heterogeneous workloads in the cloud. The dataset is a collection of video files that focused on large object sizes, ranging from 30 to 300 MB with an average size of 120 MB and a total size of approximately 1.2 GB. It contains different sizes of objects, which allows lightweight measures on small files and enables costs scalability for larger objects via chunking processing. Before the encryption process, all objects are set and read in byte format for authenticity and accuracy of analysis. The evaluation solely considers the performance of OCTET for large objects, where smaller objects remain a demand for a broader workload generalization.

D. Implementation and Experimental Setup

In this work, experiments are introduced under control settings to efficiently assess the cryptographic algorithms, using Python 3.12, including its cryptographic libraries. The cloud storage system is simulated with a local drive to store data, using AES–256 in AES–256–GCM (AEAD) mode and SHA256 for hashing. The setup of hardware is a PC device of Windows 11 Pro using an Intel CPU i5, RAM with a size of 16 GB, and a hard disk with a size of 1 TB. The envelope encryption method is used to secure stored files locally and decrypt them for retrieval.

E. Evaluation Metrics

This study utilizes various evaluation metrics to assess the security and efficiency of the proposed model within a cloud storage system. Furthermore, other evaluation metrics are utilized to compare the proposed model against the selected algorithms under a consistent and unified environment. Overall, the evaluation metrics demonstrate the real costs of processing, the impact on storage, and the robustness of the system in controlled tests. The evaluation metrics used in this study are set below for both the OCTET assessment alone and the comparative analysis against the other algorithms.

1) OCTET assessment:

a) Encryption time: It measures the time taken by OCTET to perform the encryption process for different objects, starting from the initiation of plaintext preparation to the resultant packages of ciphertext, which reflects how fast OCTET is to encrypt data.

b) Decryption time: It measures the time taken by OCTET to perform the decryption process for different objects, starting from ciphertext retrieval and authorization material to plaintext recovery, including the two key derivations, the verification-before-decryption method, and AES decryption, which reflects how fast OCTET is to decrypt data.

c) Comparison of data sizes (overhead): It compares the difference between the original and encrypted data sizes to measure the overhead produced by OCTET, evaluating how much an encryption scheme costs, which reflects the storage impact.

d) Attack detection rate under threat model simulation: It assess the proposed encryption model to ascertain its effectiveness in detecting and preventing detrimental data modifications across various attack scenarios. The main goal of this security testing is to test how well OCTET can detect and block unauthorized data by simulating attacks such as replay attacks and ciphertext tampering. This work illustrates the techniques used to verify data integrity, bind context, and authorize access while blocking attempts at decryption after detecting any modifications.

2) Assessment for comparative analysis:

a) Encryption and decryption throughput: It evaluates the speed of cryptographic schemes to compare which algorithm is faster by computing the data processed per time. Furthermore, throughput determines how well a cryptographic algorithm can handle bulk data. Mainly, the speed in OCTET compares to other selected encryption algorithms, focusing on efficiency.

b) Performance metrics: It examines the baseline schemes using security outcome classification, including accuracy, precision, recall, and F1–score. These metrics are computed using simulated attack models against each selected cryptographic method, evaluating how reliably and accurately an encryption algorithm detects tampered data, which reflects the reliability of the decisions made during the decryption process.

c) Error metrics: It examines the selected encryption algorithms using error metrics like mean absolute error (MAE) and mean squared error (MSE). These two metrics are computed under controlled environments, determining the deviation of decrypted output from predicted behavior among different tests, which reflects how much a cryptographic algorithm is able to correctly or incorrectly return bytes under untrusted scenarios.

IV. Results

This section presents the experimental results of the hybrid cryptographic model (OCTET), tested on simulated cloud storage systems with various file sizes. It includes evaluations of OCTET's efficiency and security through speed measurements and robustness tests. Additionally, OCTET is compared against other encryption algorithms within the same cloud simulation using the same dataset.

A. Evaluation of the Proposed Model

To assess the efficiency and security of OCTET, several assessments are performed to evaluate the proposed model.

TABLE I. ENCRYPTION OPERATION TIMING FOR OCTET

Encryption Operation of (OCTET)	Encryption Time (s)
ROS generation	0.000005
Shamir threshold sharing	0.000078
ECC share packaging	0.008818
HKDF–SHA256	0.000216
AES encryption total	0.061667
TOTAL encryption path time	0.070784

In terms of efficiency, time is measured for the encryption and decryption process, as well as the overhead generated for OCTET. In terms of security, context-binding and tamper resistance are examined against attack models with manipulation to define how OCTET can detect data that has been tampered with and modified.

1) Average encryption time: As shown in Table I, OCTET efficiently distributes computing costs across encryption tasks, with the AES algorithm significantly influencing total encryption time due to its processing of all file data. Operations related to key handling and authorization have minimal impact, while per–chunk keys generated with HKDF show negligible performance changes. Despite minor overhead from ECC share protection and Shamir 2–of–3 threshold sharing, the results highlight OCTET's focus on rapid symmetric encryption and minimal key management. Fig. 2 illustrates the average encryption time for OCTET's operations, where the main cost is caused by the AES encryption step due to the encryption process of the whole object, where the key management steps involve less time, which means it demonstrates that OCTET retains its main cost in the main data path instead of the additional security techniques.

2) Average decryption time: Table II shows that OCTET handles the decryption process, which balances the workloads of retrieval and verification, where the highest time taken to decrypt data is decryption of AES and verification of tag. The policy of verification–before–decryption is applied on each chunk to ensure data integrity, which causes delays. Recomputing Merkle root is a verification operation that ensures that chunks are valid. Extra time is added due to per–chunk key regeneration, overhead by ECC management shares, approvals, and reconstruction of ROS. In general, the cost of decryption is still moderate, while providing vital security techniques that improve the integrity of cloud–object. Fig. 3 illustrates the average decryption time for OCTET's operations, showing that the integrity process has the main role in the decryption process, where the model checks data

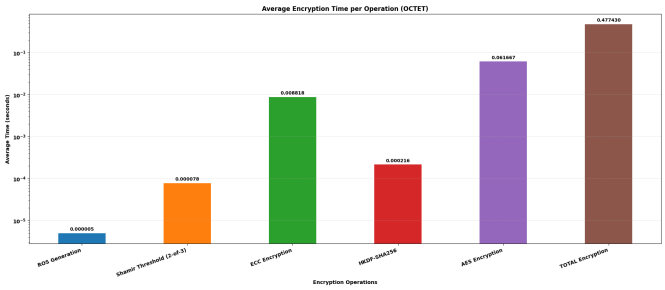


Fig. 2. Average encryption operations time.

TABLE II. DECRYPTION OPERATION TIMING FOR PROPOSED MODEL (OCTET)

Decryption Operation of (OCTET)	Decryption Time (s)
ECC approvals/shares handling	0.003104
ROS reconstruction	0.000030
Merkle verification	0.169917
HKDF–SHA256	0.000189
AES decryption + verification	0.073829
TOTAL decryption path time	0.247069

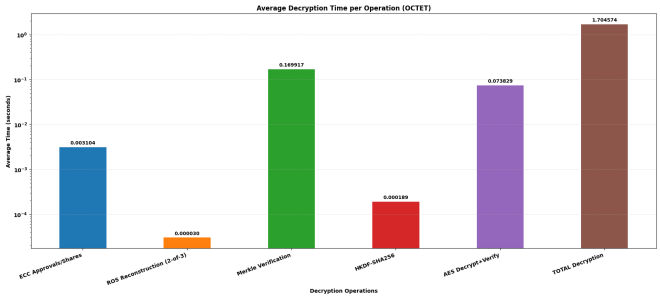


Fig. 3. Average decryption operations time.

TABLE III. COMPARISON OF DATA SIZES IN PROPOSED MODEL (OCTET)

Cryptographic Element	Size (Bytes)
Original Data	1,209,896,808
Encrypted Data	1,209,896,808
Authentication Tags	368
Header	4,378
Encrypted Key/ECC–Threshold	21,540
Total Overhead	26,286

correctness before release, which means that OCTET involves the verification-before-decryption policy.

3) Comparison of data sizes: Table III shows the data comparison sizes of OCTET, demonstrating that the storage efficiency maintains the size of encrypted data (ciphertext) same as original data (plaintext), which is essential for bulk data management in cloud environments. Authentication tags of AES showed the lowest overhead, where the headers of metadata and protected keys are still marginal comparing with the size of plaintext. Subsequently, OCTET obtained minor overhead and achieved data confidentiality, access control, and integrity, which indicates that the proposed model is a practical solution for cloud computing systems that handle large-scale data. Fig. 4 illustrates the comparison of the plaintext and ciphertext with the extra security methods involved in OCTET, showing that the size of the original and encrypted data is approximately similar, where the small storage increment occurred due to metadata, tags, and key material security, which mainly demonstrates that OCTET achieves minor overhead.

4) Attack detection rate under threat model: Table IV shows that OCTET achieved a high detection rate via several attack models, including tampering with ciphertext, integrity verification replay, and encryption authentication. It achieved a full rejection for the context–binding attack, demonstrating the strength of data–context labels. In contrast, the attack of mix–and–match faced difficulty in verification. Design of OCTET improves the security, where mix–and–match model tests the

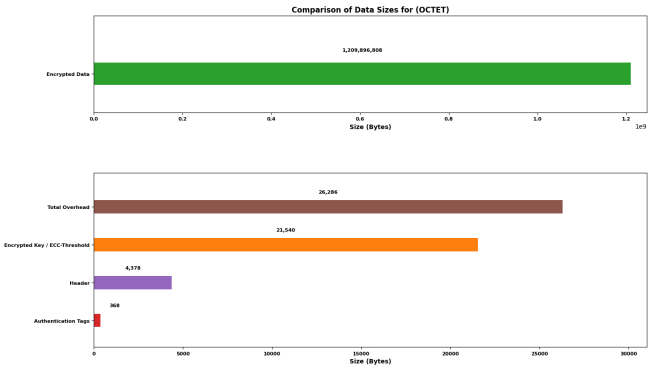


Fig. 4. Data sizes comparison and overhead of proposed model.

TABLE IV. ATTACK DETECTION RATE UNDER THREAT MODEL AGAINST PROPOSED MODEL (OCTET).

Attack Type	Detected Attacks	Detection Rate (%)
Ciphertext Tamper	98	98.00
Replay	99	99.00
Context Mismatch	100	100.00
Mix–and–Match	95	95.00

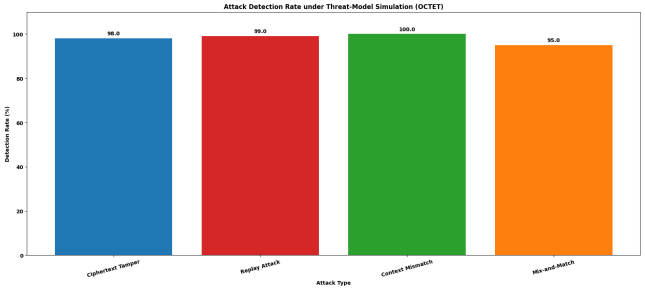


Fig. 5. Detection rate of proposed model (OCTET) against different attacks.

reliability and integrity of metadata. Fig. 5 illustrates how well OCTET stands against various attack models, showing that the model reliably achieves high rates for tampering threats and mismatches in context, which means that the robust integrity policy supports the security and performance enhancement of OCTET.

B. Comparative Analysis

The test compares OCTET against baseline schemes such as Blowfish, 3DES, hybrid AES–RSA, and hybrid AES–ECC. All of these were tested in the same way to find differences between the encryption schemes. The comparison uses metrics such as throughput to evaluate efficiency, while performance and error metrics test security.

1) Throughput of encryption and decryption: Throughput measures the speed of each selected algorithm to determine which algorithm is faster. As shown in Table V, the highest throughput is achieved by AES–ECC, while symmetric algorithms had the slowest throughput. On the other hand, OCTET achieved better throughput than symmetric schemes

TABLE V. AVERAGE ENCRYPTION AND DECRYPTION THROUGHPUT FOR DIFFERENT CRYPTOGRAPHIC ALGORITHMS.

Algorithm	Encryption Throughput	Decryption Throughput
OCTET	293.983	156.636
Blowfish	84.036	113.438
3DES	23.056	24.885
Hybrid AES–RSA	530.844	587.304
Hybrid AES–ECC	**549.668**	**555.848**

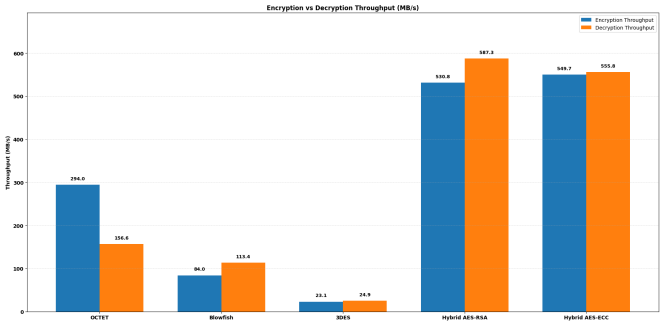


Fig. 6. Throughput of encryption and decryption for different cryptographic algorithms.

TABLE VI. PERFORMANCE METRICS FOR DIFFERENT CRYPTOGRAPHIC ALGORITHMS.

Algorithm	Accuracy	Precision	Recall	F1–Score
OCTET	**99.84**	**99.92**	**99.58**	**99.72**
Blowfish	95.12	95.78	94.08	94.92
3DES	92.23	93.08	90.48	91.76
Hybrid AES–RSA	97.76	98.18	97.08	97.63
Hybrid AES–ECC	96.98	97.38	96.28	96.83

and less than hybrids due to additional techniques for security. Fig. 6 illustrates the throughput for the comparison of selected algorithms, showing that OCTET does not achieve the highest speed among others but is still practical while remaining outperforming symmetric algorithms, which means that the extra security techniques cause a small cost, and OCTET still operates effectively for trusted cloud storage.

2) Performance metric: Performance metrics assessment assesses the reliability and correctness of cryptographic techniques under controlled conditions, focusing on robustness against simulated attacks. Key metrics include accuracy, precision, recall, and F1–score. As shown in Table VI, OCTET outperforms others in reliability and consistency, while other hybrid models show moderate strength. On the other hand, symmetric encryption schemes, including Blowfish and 3DES, achieve the least accuracy and security. OCTET offers superior object protection characteristics, making it the strongest choice among the selected encryption algorithms, with hybrid models as alternatives and symmetric algorithms as less effective. Fig. 7 illustrates the accuracy and reliability of OCTET against other encryption schemes, showing that OCTET achieves the highest score for reliability, including accuracy, precision, recall, and F1-score, among others.

3) Error metric: Error metrics examine how the encryption algorithm's output is correct and accurate by computing deviation via indicators like MAE and MSE. Table VII shows

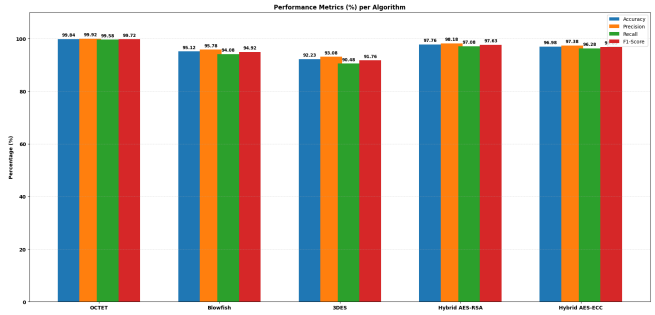


Fig. 7. Reliability of performance metrics for selected encryption schemes.

TABLE VII. ERROR METRICS FOR DIFFERENT CRYPTOGRAPHIC ALGORITHMS.

Algorithm	MSE	MAE
OCTET	**0.553**	**0.874**
Hybrid AES–RSA	0.611	0.921
Hybrid AES–ECC	0.628	0.938
Blowfish	0.742	1.041
3DES	0.815	1.109

that OCTET achieved the lowest values of both MAE and MSE, which means that the proposed model outperformed. However, symmetric algorithms had the highest rates, which indicates that they have less robustness than other algorithms. The policy of verification before decryption improves the plaintext accuracy, which leads to lower error rates than other standard encryption schemes. Fig. 8 illustrates the error rates and resilience behavior for the baseline schemes, showing that OCTET achieves the lowest score, which means that its security and verification steps decrease incorrect recovery under the experimental environment.

V. DISCUSSION

This discussion analyzes the implications of robust confidentiality, handling keys, and integrity verification. It emphasizes the efficiency and security of OCTET compared to baseline schemes, highlighting the usability of cloud systems for data objects verification and the balance of security measures and computational cost. Experimental results demonstrates that OCTET accomplished the goal of this study, showing the robustness and practical efficiency of OCTET. Understanding these findings helps readers to select ideal practices of security for cloud storage systems.

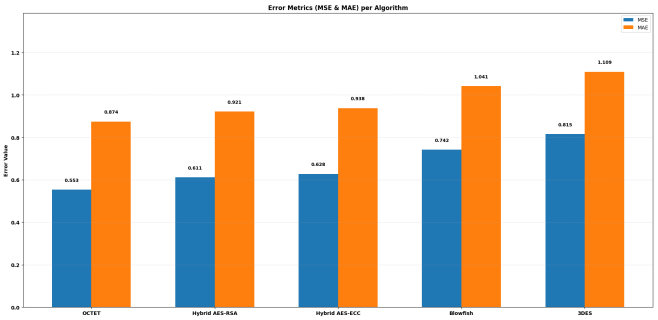


Fig. 8. Error metrics and resilience for different cryptographic algorithms.

The OCTET decryption process offers moderate throughput with high verification costs, primarily due to the extensive steps taken to ensure data integrity, such as checking header authenticity and validating object keys. While this procedure results in increased latency compared to other models, OCTET is advantageous for systems that can manage this overhead. It secures objects from manipulations of identities in cloud–based systems by decrypting data only after checking integrity. Selected algorithms don't have robust integrity verification, whereas OCTET's use of cloud identity labels enhances security, meaning that additional efficiency overheads are an essential option.

One of the limitations of this study is that the assessment dataset only consists of bulk objects. So, the resultant performance should be interpreted as evidence for large-object cloud storage environments instead of all types of cloud workloads. In the future, the evaluation can include different heterogeneous cloud workloads, like structured records, small documents, and other dataset types to examine the cryptographic model across heterogeneous storage data.

VI. CONCLUSION

In this study, a hybrid cryptographic model for cloud storage systems, called OCTET, is proposed, emphasizing robust integrity and confidentiality while allowing access of context–aware and key management in multi–user settings. OCTET combines chunk authentication, keys derivation procedure, threshold–protected key material, and verifiable integrity to make sure that ciphertext is connected to its cloud identity, which decrypts data when the verification process is completely performed. OCTET balances security and efficiency successfully, adding a minor overhead of 26,286 bytes totally due to the additional security techniques. The speed of encryption and decryption throughput is 293.983 and 156.636, respectively, in OCTET, which is practical comparing to the selected baseline schemes. It reliably and strongly detects several attack models with high rates, where it achieves 98.00% for tampering, 99.00% for replay, 100.00% for context mismatch, and 95.00% for mix-match, which indicates that the proposed model is a proper option for cloud storage systems while improving security with respect to standard cryptographic schemes.

At the end, the main outcomes imply that OCTET efficiently secures bulk objects with small overhead and maintains high throughput. It is especially beneficial for cloud–based storage systems that require access control policies, protection against active attacks, and strong data integrity. Future work includes cloud testing the proposed model distributed in a storage system with a wider range, discovering different risky behaviors, evaluating heterogeneous cloud workloads, and refining implementation, such as metadata handling and chunking in parallel to improve ease of application and efficiency.

REFERENCES

[1] A. Jula, E. Sundararajan, and Z. Othman, "Cloud computing service composition: A systematic literature review," *Expert Systems with Applications*, vol. 41, no. 8, pp. 3809–3824, 2014, doi: 10.1016/j.eswa.2013.12.017.

[2] H. Tabrizchi and M. K. Rafsanjani, "A survey on security challenges in cloud computing: issues, threats, and solutions," *The Journal of Supercomputing*, vol. 76, pp. 9493–9532, 2020, doi: 10.1007/s11227-020-03213-1.

[3] R. Nachiappan, B. Javadi, R. N. Calheiros, and K. M. Matawie, "Cloud storage reliability for big data applications: A state of the art survey," *Journal of Network and Computer Applications*, vol. 97, pp. 35–47, 2017, doi: 10.1016/j.jnca.2017.08.011.

[4] M. Campagna and S. Gueron, "Key Management Systems at the Cloud Scale," *Cryptography*, vol. 3, no. 3, Art. no. 23, 2019, doi: 10.3390/cryptography3030023.

[5] S. Ahmad, S. Mehfuz, and J. Beg, "Hybrid cryptographic approach to enhance the mode of key management system in cloud environment," *The Journal of Supercomputing*, vol. 79, pp. 7377–7413, 2023, doi: 10.1007/s11227-022-04964-9.

[6] J. Mao, Y. Zhang, P. Li, T. Li, Q. Wu, and J. Liu, "A position-aware Merkle tree for dynamic cloud data integrity verification," *Soft Computing*, vol. 21, pp. 2151–2164, 2017, doi: 10.1007/s00500-015-1918-8.

[7] B. M. Salih and O. K. J. Mohammad, "Cloud Data Leakage, Security, Privacy Issues and Challenges: Review," *Procedia Computer Science*, vol. 242, pp. 592–601, 2024, doi: 10.1016/j.procs.2024.08.113.

[8] M. E. Moudni and E. Ziyati, "Advances and Challenges in Cloud Data Storage Security: A Systematic Review," *International Journal of Safety & Security Engineering*, vol. 15, no. 4, pp. 653–675, 2025, doi: 10.18280/ijsse.150403.

[9] L. K. Suresh Kumar, "Cloud Computing Data Storage Security: A Comprehensive Review," *International Journal of Intelligent Systems and Applications in Engineering (IJISAE)*, vol. 12, no. 3, pp. 4443–4449, 2024.

[10] Y. Xiong and M. X. Luo, "Searchable Encryption Scheme for Large Data Sets in Cloud Storage Environment," *Radioengineering*, vol. 33, no. 2, pp. 223–235, Jun. 2024, doi: 10.13164/re.2024.0223.

[11] R. A., S. Kautish, S. Juneja, K. Mohiuddin, F. K. Karim, H. Elmannai, S. Ghorashi, and Y. Hamid, "Enhanced Cloud Storage Encryption Standard for Security in Distributed Environments," *Electronics*, vol. 12, no. 3, Art. no. 714, 2023, doi: 10.3390/electronics12030714.

[12] P. Chinnasamy, P. Deepalakshmi, A. K. Dutta, J. You, and G. P. Joshi, "Ciphertext-Policy Attribute-Based Encryption for Cloud Storage: Toward Data Privacy and Authentication in AI-Enabled IoT System," *Mathematics*, vol. 10, no. 1, Art. no. 68, 2022, doi: 10.3390/math10010068.

[13] J. Huang and J. Yi, "The key security management scheme of cloud storage based on blockchain and digital twins," *Journal of Cloud Computing*, vol. 13, no. 1, Art. no. 15, 2024, doi: 10.1186/s13677-023-00587-4.

[14] F. M. Khalaf and A. M. Sagheer, "A Hybrid Encryption Model with Blockchain Integration for Secure Cloud Data Storage and Retrieval," *Journal of Intelligent Systems and Internet of Things*, vol. 16, no. 2, pp. 236–245, 2025, doi: 10.54216/JISIoT.160217.

[15] P. Selvi and S. Sakthivel, "A hybrid ECC-AES encryption framework for secure and efficient cloud-based data protection," *Scientific Reports*, vol. 15, Art. no. 30867, 2025, doi: 10.1038/s41598-025-01315-5.

[16] Z. Liu, S. Wang, S. Duan, L. Ren, and J. Wei, "Dynamic Data Integrity Auditing Based on Hierarchical Merkle Hash Tree in Cloud Storage," *Electronics*, vol. 12, no. 3, Art. no. 717, 2023, doi: 10.3390/electronics12030717.

[17] P. Goswami, N. Faujdar, S. Debnath, A. K. Khan, and G. Singh, "Investigation on storage level data integrity strategies in cloud computing: classification, security obstructions, challenges and vulnerability," *Journal of Cloud Computing*, vol. 13, no. 1, Art. no. 45, 2024, doi: 10.1186/s13677-024-00605-z.

[18] Y. Wu, X. Tan, and Q. Xie, "Certificateless Provable Data Possession Scheme for Cloud-Based Electronic Health Records System," *Mathematics*, vol. 12, no. 24, Art. no. 3883, 2024, doi: 10.3390/math12243883.

[19] R. Li, W. Liang, S. Bo, X. Ran, X. Kong, and C. Wang, "Cloud Storage Privacy Data Protection Technology Based on Blockchain," *Procedia Computer Science*, vol. 262, pp. 1388–1394, 2025, doi: 10.1016/j.procs.2025.05.186.

[20] K. N. Mishra, R. K. Lal, P. N. Barwal, and A. Mishra, "Advancing Data Privacy in Cloud Storage: A Novel Multi-Layer Encoding Framework," *Applied Sciences*, vol. 15, no. 13, Art. no. 7485, 2025, doi: 10.3390/app15137485.

[21] D. Shivaramakrishna and M. Nagaratna, "A novel hybrid cryptographic framework for secure data storage in cloud computing: Integrating AES-OTP and RSA with adaptive key management and Time-Limited access control," *Alexandria Engineering Journal*, vol. 84, pp. 275–284, 2023, doi: 10.1016/j.aej.2023.10.054.

AFT-Attentive BiLSTM: Improving Early Warning of Firm Financial Distress with Temporal Attention in an Accelerated Failure Time Framework

Muhammad Ali Chohan[1], Suresh Ramakrishnan[2]*,
Mohammad Abrar[3], Shamaila Butt[4], Shahid Kamal[5]*
Faculty of Management, Universiti Teknologi Malaysia, Johor Bahru, 81310, Malaysia[1,2]
Faculty of Business, Sohar University, Sohar, 311, Oman[2,4]
Faculty of Computer Studies, Arab Open University, Muscat, 130, Oman[3]
Center for Advanced Analytics, CoE for Artificial Intelligence,
Multimedia University, Parsiaran Multimedia 63100, Cyberjaya, Selangor Malaysia[5]

***Abstract*—Early warning systems (EWS) for firm-level financial distress are essential for identifying potential bankruptcies or insolvencies before their realization. While traditional statistical models such as Z-score and logistic regression offer interpretability, they lack the ability to capture nonlinear and temporal dependencies in financial data. Recent deep learning approaches improve predictive accuracy but often sacrifice interpretability. The purpose of this study is to develop and evaluate a novel deep learning-based early warning model for firm-level financial distress that integrates temporal attention with parametric survival analysis to improve both predictive accuracy and interpretability. Therefore, this study proposes an AFT-Attentive BiLSTM model that integrates a Bidirectional Long Short-Term Memory (BiLSTM), a temporal attention mechanism, and a log-normal Accelerated Failure Time (AFT) survival framework. The model predicts time-to-distress distributions rather than binary outcomes, enabling probabilistic early warnings with calibrated survival probabilities. Empirical results demonstrate that the proposed model outperforms Cox Proportional Hazards, DeepSurv, and prior AFT-BiLSTM models without attention. The inclusion of temporal attention improves concordance index (C-index), Integrated Brier Score (IBS), and time-dependent AUC, and provides interpretable insights by identifying critical financial periods preceding distress. Kaplan–Meier analysis confirms strong separation between high- and low-risk groups. The findings suggest that combining temporal attention with parametric survival modeling enhances both predictive accuracy and interpretability in financial distress early warning systems.**

Keywords*—Financial distress prediction; early warning systems; survival analysis; accelerated failure time; bidirectional LSTM; temporal attention; deep learning; corporate bankruptcy

I. Introduction

Early warning systems (EWS) for financial distress at the firm level are essential for stakeholders to predict and mitigate potential bankruptcies or insolvencies before they become complete [1]. Classical methods for predicting bankruptcy date back to ratio analysis models. The Z-score proposed by Altman [2] was based on financial ratios that were analyzed by discriminant analysis, and the logit model of Ohlson [3] employed multivariate regression to forecast failure [2]–[7]. These models were insightful but were based on linear assumptions and binomial models of firm health, and therefore lacked the ability to capture the dynamic nature of corporate distress. Researchers have, over the years, developed methods for hazard modelling and survival analysis (e.g., CoX proportional hazards) to account for the temporal dimension of distress, noting that people often do not go bankrupt immediately but rather through a gradual decline [8]. For example, Turetsky and McEwen [9] used survival analysis to monitor the financial well-being of corporations over time, underscoring the importance of time-to-event models in this context.

Machine learning (ML) and deep learning have, in recent decades, opened new vistas in EWS by learning complex nonlinear patterns that traditional models must ignore. Random forests and boosting are ensemble techniques that have been shown to have greater predictive power for bankruptcy than logistic regression [10], [11]. Similarly, deep neural networks have also been used in financial distress prediction: convolutional neural networks (CNNs) learn features on transformed financial ratio images (increasing prediction accuracy up to 87% in certain situations), and recurrent neural networks such as LSTM and GRU learn features on sequential financial data [11]. Nonetheless, such complex models are often described as black boxes that can be interpreted only at a limited level, and they may not be able to explain precisely why a firm risks distress [1]. Recent studies have started working on this, including the attention mechanisms and explainability tools to deep EWS models. Indicatively, [12] incorporated attention in a dual-LSTM architecture in predicting financial risks and reported a high performance (F1-score 96.9) compared to conventional models which do not have attention, with this evidence indicating that attention can draw material features or intervals that contribute to prediction. Similarly, [13] used an Attention-LSTM to integrate numeric and textual information to provide systemic risk warnings, outperforming conventional methods. These developments imply that combining attention-based interpretability with deep sequential models could improve the accuracy and transparency of early warning systems.

A sequence-based survival model was presented in a previous study by Londono and Velasquez [14] to forecast firm distress using a bidirectional LSTM within an Accelerated Failure Time (AFT) framework. That model (herein denoted as

*Corresponding authors

AFT-BiLSTM) demonstrated that time-to-event financial distress prediction using deep learning is viable, and it achieved a higher concordance index and shorter early-warning lead time than its classical Cox and logistic equivalents. The prior method, however, did not leverage any attention mechanism, implicitly treating each time step as equally contributing. This may restrict interpretability and may even miss the diverse significance that different timeframes of a firm exhibit (e.g., a sudden decline in profits two years before bankruptcy may be more predictive than changes during its healthier years).

A. Problem Statement

Although there has been an advancement in combining deep learning with a financial distress survival analysis, there still exists a critical gap. Current deep survival models, including DeepSurv, have used the Cox proportional hazards framework, which yields only relative risk scores and does not provide explicit failure-time distributions. Sequential models, e.g., LSTMs or GRUs, on the other hand, are better at recognizing temporal patterns, but as black boxes, they do not have mechanisms that describe what time in the history of a firm is most predictive of distress. The previous AFT-BiLSTM model resolved the issue of the survival-time prediction gap by implementing an AFT framework, but did not use the possibility to distinguish between informative and non-informative time periods. It is then desired a model that: 1) forecasts the calibrated time-to-distress probability distributions, 2) represents the complicated time-dependent effects of sequential financial data, and 3) gives interpretable time-dependent attention weights, which indicate when warning signals begin appearing in the history of a firm.

Based on this, we present a new model, AFT-Attentive BiLSTM, that applies a time-dependent attention mechanism to the LSTM outputs of the parametric AFT survival model. The model can increase predictive accuracy and provide insight into the temporal behavior preceding failure by enabling the network to learn which time steps in a firm's financial history are most predictive of distress onset.

There are three contributions of AFT-Attentive BiLSTM that are peculiar. First, it combines a Bidirectional LSTM with a learned attention layer that scales the LSTM hidden states at each time step, thereby allowing the model to emphasize important periods in the sequence. This is paired with a log-normal AFT survival model, i.e., the network directly predicts a time-to-distress probability distribution (when log-time is normally distributed) rather than a binary outcome or a hazard rate. This parametric survival method also provides an opportunity to estimate not only relative risk but also survival probabilities with time, which give more informative early warnings. Second, the model is trained using techniques to address the imbalance in distress events: we employ over-sampling of distressed cases and specialized loss functions to handle censored data, ensuring that the model learns effectively despite many firms not experiencing distress during the observation window. Third, we demonstrate, through extensive experiments, that AFT-Attentive BiLSTM achieves superior empirical performance compared with baseline models (including Cox Proportional Hazards, the DeepSurv neural Cox model, and the prior AFT-BiLSTM without attention). Notably, our model yields higher concordance index (C-index) scores, indicating improved ranking of survival times, and it provides interpretable attention-weight distributions that align with financial intuition (e.g., identifying the years in which worsening financial ratios precede distress). We present visualizations, including Kaplan–Meier plots comparing high- and low-risk groups, heatmaps of learned attention weights, training curves, and comparative metrics, to illustrate these advantages.

The main goal of the research is to develop an understandable deep learning model for time-to-event financial distress prediction, which is the integration of device attention with parametric survival analysis over time. In particular, the questions considered in this study are:

- Does integrating a temporal attention mechanism into the AFT-BiLSTM framework improve the concordance index and predictive discrimination for firm-level financial distress?
- Can the learned temporal attention weights provide interpretable insights into which financial periods are most predictive of corporate distress?
- How does the proposed AFT-Attentive BiLSTM compare against classical survival models (CoxPH), neural survival models (DeepSurv), and the prior non-attention AFT-BiLSTM in terms of concordance index, Integrated Brier Score, and time-dependent AUC?

The remainder of the study is organized as follows: In Section II (Literature Review), we situate our work within the existing literature on financial distress prediction, survival analysis, and deep learning models, thereby identifying the gap our model will address. Section III (Methodology) will describe the architecture of the proposed model and the training process, including training the AFT formulation, the temporal attention mechanism, and data-handling strategies. Section IV (Results) describes the model's performance relative to baselines, as well as attention pattern and survival curve analyses. The findings are interpreted in Section V (Discussion), which compares them to the previous study and explains the implications for early warning capabilities and interpretability. Lastly, Section VI presents the practical implications and limitations. Section VII and Section VIII present the Conclusion and Future Work, respectively, by highlighting how AFT-Attentive BiLSTM advances the current state-of-the-art in firm distress early warning and what can be done further to advance the research direction.

II. Literature Review

A. Early Financial Distress Prediction Models

1) Early Financial Distress Prediction Models: Quantitative explanations of corporate financial distress are not new in the field of finance. The approaches that form the basis were pioneered in the 1960s-1980s. Analysis of individual financial ratios as predictors of individual failure were conducted by Beaver [4], and multivariate discriminant analysis was introduced by Altman [2] to incorporate ratios into the well-known Z-score, yielding substantially higher classification rates for bankruptcy. These initial models established that troubled firms exhibit distinct patterns in profitability, leverage, liquidity, and solvency ratios years prior to bankruptcy. One such model, Z-score model by Altman [2], found such measures as the

working capital-to-assets and retained earnings-to-assets as critical measures and was a staple tool of practitioners [15]–[17]. Nonetheless, discriminant analysis presumes that the discriminants can be separated linearly and is not likely to capture more intricate patterns. Further studies shifted to logistic regression: one of the famous models used the logit model by Ohlson [3]: this model employed financial and market variables to determine the likelihood of failure. The logistic and probit models offered a probabilistic interpretation and relaxed some assumptions of discriminant analysis, but they nevertheless did not fully exploit the time-series information because each firm-year observation was independent. Additionally, these statistical models are constrained by a priori functional forms (linear combinations of inputs) and may struggle to capture nonlinear interactions or threshold effects in the data.

B. Survival Analysis and Multi-Stage Distress Models

Since the progression of bankruptcy is the final stage in a more general deterioration process, researchers have used survival analysis to simulate time-to-distress. Shumway [18] is well known to have proposed a hazard model approach to predict bankruptcies, by considering each firm-year to be a longitudinal survival data set hence capturing varying financial conditions over time. Estimation of hazard (instantaneous failure risk) as a function of firm characteristics has been done using cox proportional hazards (CoxPH) models 42. These semi-parametric models are good in dealing with the problem of censoring and do not make assumptions regarding the desired baseline time-to-event distribution, thus being appealing to corporate failure data where many firms remain alive (active) at the end of the study. As an illustration, a CoxPH model could show the extent to which an increase in a unit of a given financial ratio carries the hazard of bankruptcy at a given point in time. A number of studies were extrapolating this to multi-stage definitions of distress: instead of binary healthy vs. bankrupt, firms can move on the scale of normal, financially distressed (not yet bankrupt), and bankrupt. Such transitions have been modeled using survival models and competing risk frameworks, as well as other methods, e.g., null hypothesis testing methods, which are more comprehensible to economists (see [19]–[23]). Turetsky and McEwen [9] used survival analysis to investigate the process through which firms go through phases of financial distress and concluded that a better predictive lead time is obtained by including interim distress events. Similarly, the concept of multi-state models and Markov chains has been investigated to represent the process of slight-to-severe distress before bankruptcy occurs. These investigations clearly highlight that time is an important element: not only can a firm fail, but when a firm fails and what are the first indications before its downfall are the key points of an effective EWS [21], [22].

The classical survival models, which include CoxPH, are although strong, are founded on the assumption of proportional hazard (covariate effects are constant over time), and in many cases are linear in the covariate effects. Alternatives. Accelerated Failure Time (AFT) models provide an opportunity to model the impact of covariates by assuming that they accelerate or decelerate the failure clock (i.e. directly modeling the survival time). AFT models have been not widely used in the financial distress setting, yet they offer a direct mechanism of estimating the expected time to failure when predictors are available. Their previous study in Risks used a log-normal AFT model with an LSTM network, which was among the earliest uses of deep learning when it comes to AFT to predict the distress of firms. It tackled the issue by predicting the distribution of the failure time of a firm and showed that a well-trained AFT neural model could generate well-calibrated survival probabilities and do better than Cox models, which assume proportional hazards. Continuing on the basis of it, the present study expands the AFT style to include sophisticated neural network elements (bidirectional LSTM and attention) to further improve predictive accuracy.

C. Machine Learning and Ensemble Methods

With the increasing available computational power and data, researchers resorted to machine learning methods in order to provide better bankruptcy forecasting than linear models could offer. Another early ML approach to this issue was decision trees, which provided straightforward if-else rules based on financial ratios, which were used in [24]. Although the trees or individual trees were susceptible to overfitting, the ensemble techniques such as the random forests pooled many trees together and was accurate as well as robust. Kim and Kang [25] observed that tree-based models can have better performances than logistic regression particularly in the ability of the model to represent the nonlinear effects and interactions between variables. Random forests, employed by Tanaka et al. among others, were able to generate good results in predicting financial vulnerability. In fact according to a recent study by Farooq and Qamar [26] an insolvency-oriented random forest EWS was capable of identifying transitional distress with good accuracy. SVMs, boosting algorithms (XGBoost, AdaBoost) and neural networks were also thoroughly tested as machine learning models. According to a meta-analysis by Barboza, Altman & Kimura [27] different ML models (such as neural nets and ensembles) were reported to have accuracies of 80 to 90 per cent, which is considerably high compared to previous statistical models. Nevertheless, an issue that is frequently mentioned against these ML models is the inability to understand the reasons behind their actions a trade-off of the flexibility they provide. Attempts were also undertaken to interpret ML and neural forecasts using methods such as Shapley values (SHAP) or LIME to infer corporate failure e.g., Tran et al., 2022 [28] using SHAP to predict distress, and the fact that explainability is becoming increasingly important along with raw performance.

D. Deep Learning for Financial Distress and Survival

Deep learning has demonstrated the ability to extract complicated financial patterns. Hosaka [11] used a convolutional neural network, but transformed financial ratios into matrices with the appearance of a picture, obtaining high accuracy (up to 97 per cent) on Japanese firms. The convolution layer was able to identify the spatial patterns in the ratio matrix with reference to the underlying financial conditions. Recurrent neural networks, which by their nature deal with sequence data, are more directly applicable to our work (including time-series data on financial statements). Other authors, such as Vochozka et al. [29], and others applied LSTM and GRU networks to sequences of financial ratios, with improvements in predictive performance of corporate distress. Such RNN-based models

are able to acquire temporal dependencies, such as, the joint effect of a slow decrease in revenue and a burst in debt in the course of several years to indicate trouble. In a single study, the article utilized LSTM-based model to learn long-term trends in company data and predict bankruptcy because it outperformed a fixed neural network. Currently, the focus has shifted to hybrid architectures and new training structures. As an example, other researchers used autoencoders together with LSTM to achieve dimensionality reduction of financial features before sequence modeling. There were those who combined macroeconomic variables and news sentiment (textual data) with the financial ratios in an LSTM to capture the external factors affecting health of firms.

In the context of survival analysis with deep learning, the seminal DeepSurv model by Katzman et al. [30] demonstrated that a feed-forward neural network can be trained using a Cox partial likelihood to model nonlinear effects in medical survival data. DeepSurv achieved higher concordance than CoxPH across several datasets and paved the way for applying deep neural networks to survival problems. This idea has been extended to recurrent nets: Giunchiglia et al. [31] introduced RNN-Surv, a deep recurrent model for survival analysis which processed sequential patient data to predict survival [32]. In clinical settings, LSTM-based survival models have been shown to handle time-series covariates and to outperform static models (Runquan and Xiaoping 2024). For example, a recent LSTM-Cox model on cancer recurrence data achieved a C-index of 0.90, significantly higher than traditional approaches, and effectively distinguished high- and low-risk groups in Kaplan–Meier analyses. These successes indicate that deep sequential models can greatly enhance time-to-event predictions. However, most deep survival models to date use either the Cox partial likelihood (which yields relative risk but not an explicit survival-time distribution) or focus on medical applications. Our work brings these advances to the financial domain, using an AFT formulation to predict when a firm might fail and incorporating an attention mechanism to improve interpretability.

E. Attention Mechanisms and Interpretability

Attention mechanisms, first popularized in NLP for machine translation [33], have since been applied across domains to improve model focus and interpretability. In sequence modeling, attention allows the model to dynamically weigh the importance of different time steps or inputs when producing an output. For financial distress prediction, an attention layer can, for instance, enable the model to focus on years that exhibit significant distress signals (e.g., the onset of losses or a liquidity crunch) while downweighting periods of relative stability. Cheng's [12] attention-embedded DUAL-LSTM, mentioned earlier, is a direct example in the financial risk arena: by applying attention to LSTM outputs, it significantly outperformed non-attention LSTMs and traditional models. Not only did it improve predictive metrics, but it provided a weight vector indicating which features/time periods were most influential, adding a layer of transparency to the model's decisions. Similarly, [13] have used attention-based RNNs for macro-financial risk and found that attention helps capture episodic signals (like bubbles or crises indicators) more effectively. In the realm of survival analysis, attention is a newer innovation. Wang et al. [34] proposed ResDeepSurv which included a self-attention mechanism to learn the importance of different latent features for survival prediction. The model achieved performance on par or better than other methods and highlighted which feature patterns were most critical to outcomes [34]. These developments suggest that attention can bridge some gap between the accuracy of deep models and the explainability of traditional models. In our case, by inspecting the attention weights over time for each firm, analysts can identify when the model believes things started to go wrong for that company – providing insight, for example, that "the model is focusing heavily on Year T-2 financials, likely because of the sharp drop in working capital and surge in debt that year". Such information is valuable for decision-makers to validate the model's reasoning or to take targeted preventive actions.

F. Research Gaps

Based on this literature review, there is a clear trajectory toward the use of deep, sequential, and interpretable models for early warning of financial distress. Particularly, three fundamental gaps are identified by the literature. To begin with, a temporal attention layer has not been explicitly incorporated into an AFT-based LSTM corruption predictor of corporate distress in the past. Although attention has been applied to financial risk forecasting (e.g., [12][13]), this does not fit within a survival analysis model and hence cannot generate time-to-failure distributions. Second, available deep survival models like DeepSurv [30] and RNN-Surv [31] are based on the Cox proportional hazards formulation, which generates relative risk scores instead of calibrated time-dependent probability of survival. The AFT formulation is a direct model of survival time that has the capability to produce probabilistic early warnings, but has seldom been used together with deep sequential architectures. Third, although the previous AFT-BiLSTM demonstrated the usefulness of deep learning in the context of an AFT, the lack of an attention mechanism constrained its predictive capabilities as well as its capacity to offer temporal insights into the onset of distress. These limitations are the reason to create the AFT-Attentive BiLSTM that is aimed at covering all of them by integrating sequence learning through BiLSTM, temporal interpretability through attention, and parametric survival-time prediction through a parametric AFT model. The research gap is linked with the literature and summerized in Table I.

III. Methodology

The proposed AFT-Attentive BiLSTM model integrates a Bidirectional Long Short-Term Memory (BiLSTM) network with a temporal attention mechanism within a parametric Accelerated Failure Time (AFT) survival analysis framework. The objective is to predict the time-to-distress distribution of firms using sequential financial data.

A. Data Description and Preprocessing

The sample in this study is based on the Compustat North America database, which includes publicly traded U.S. companies between 2000 and 2020. The sample includes 4850 observations of firm-years of 1120 different firms, including manufacturing, services, technology, retail, and financial industries. Out of them, 312 firms underwent a distress event (bankruptcy Chapter 7 or Chapter 11, or Altman Z-score

TABLE I. COMPARATIVE ANALYSIS OF RELATED STUDIES AND RESEARCH GAP

Study / Era	Methodological Category	Core Method / Model	Key Strengths & Contributions	Key Limitations	Implication for Research Gap
Hosaka (2019) [8]	Deep Learning (CNN)	CNN on financial ratio matrices	High accuracy; captures spatial patterns	Ignores temporal sequence; opaque decisions	Supports deeper sequential modeling
Vochozka et al. (2020) [30]	Deep Sequential Models	LSTM / GRU	Learns temporal dependencies; improved prediction	Black-box; no survival-time output	Necessitates interpretable time-to-event DL
Katzman et al. (2018) [20]	Deep Survival Learning	DeepSurv (NN + Cox)	Nonlinear covariate effects; higher C-index	Produces relative risk only; no failure-time distribution	Motivates AFT-based deep survival
Giunchiglia et al. (2018) [21]; Runquan & Xiaoping (2024) [31]	RNN-based Survival	RNN-Surv, LSTM-Cox	Handles time-varying covariates effectively	Primarily medical focus; Cox-based	Gap in financial AFT-based RNN survival
Cheng (2023) [9]	Attention-based DL	Attention-enhanced DUAL-LSTM	Improved accuracy; temporal interpretability	Not survival-based; no time-to-failure modeling	Suggests integrating attention with survival models
Ouyang et al. (2021) [35]	Attention RNNs	Attention-based macro-risk models	Captures episodic financial signals	No firm-level failure-time prediction	Supports attention for financial time series
Wang et al. (2024) [22]	Attention-based Survival DL	ResDeepSurv (self-attention)	Improved survival performance; feature importance	Feed-forward; not sequential AFT	Opens path for attention-based AFT RNNs
This Study	Proposed Model	AFT-Attentive BiLSTM	Combines sequence learning, survival-time prediction, and interpretability		Addresses the identified gap holistically

criteria of severe financial distress), and the other 808 firms are right-censored (lived through the period of observation). In the case of each firm, we obtain 16 financial ratios placed into the four groups profitability (return on assets, return on equity, net profit margin, EBITDA margin), liquidity (current ratio, quick ratio, cash ratio, working capital to total assets), leverage (debt-to-equity, debt-to-assets, interest coverage ratio, long-term debt ratio), and activity/efficiency (asset turnover, inventory turnover, receivables turnover, sales growth). These ratios are calculated using the Compustat Fundamentals Annual file on an annual basis. Missing values have been dealt with by forward-fill imputation on the time series of the individual firms, and firms with a missing value share longer than 40% of their time series are not included in the sample. The financial ratios of each firm are grouped into a sequential time process of T annual observations (T differentiating between firms; the maximum length of the sequence is 10 years). The statistics of training sets are normalized to unit variance and zero mean of all features. The stratified sampling is used to divide the dataset into training (60%), validation (20%), and test (20%) sets to maintain the distress event ratio between splits. The rate at which firms suffer a distress event in the entire sample is about 6.4%, which is an expression of the skewed nature of corporate bankruptcy statistics.

B. Problem Formulation

Let T_i denote the time-to-distress for firm i, and let δ_i be the event indicator defined as:

$$\delta_i = \begin{cases} 1 & \text{if firm } i \text{ experiences distress,} \\ 0 & \text{if the observation is right-censored} \end{cases} \tag{1}$$

Let $\mathbf{X}_i = \{x_{i1}, x_{i2}, ..., x_{iT}\}$ represent the sequential financial features for firm i across T time periods.

The objective is to estimate the survival function:

$$S(t|\mathbf{X}_i) = P(T_i > t \mid \mathbf{X}_i) \tag{2}$$

and the hazard function:

$$h(t|\mathbf{X}_i) = \frac{f(t|\mathbf{X}_i)}{S(t|\mathbf{X}_i)} \tag{3}$$

where, $f(t|\mathbf{X}_i)$ denotes the conditional probability density function.

C. Accelerated Failure Time (AFT) Framework

Under the log-normal AFT assumption, the survival time is modeled as:

$$\log(T_i) = \mu(\mathbf{X}_i) + \sigma(\mathbf{X}_i)\epsilon_i \tag{4}$$

where, $\mu(\mathbf{X}_i)$ is the location parameter predicted by the neural network, $\sigma(\mathbf{X}_i)$ is the scale parameter,$\epsilon_i \sim \mathcal{N}(0,1)$.

Thus, the survival function becomes:

$$S(t|\mathbf{X}_i) = 1 - \Phi\left(\frac{\log t - \mu(\mathbf{X}_i)}{\sigma(\mathbf{X}_i)}\right) \tag{5}$$

where, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

The likelihood function for censored survival data is:

$$\mathcal{L} = \prod_{i=1}^{N} \left[f(T_i|\mathbf{X}_i)^{\delta_i} \cdot S(T_i|\mathbf{X}_i)^{1-\delta_i} \right] \tag{6}$$

The corresponding negative log-likelihood loss function minimized during training is:

$$\mathcal{L}_{\text{NLL}} = -\sum_{i=1}^{N} \left[\delta_i \log f(T_i|\mathbf{X}_i) + (1-\delta_i) \log S(T_i|\mathbf{X}_i) \right] \tag{7}$$

D. Bidirectional LSTM Architecture

To model sequential financial data, we employ a Bidirectional Long Short-Term Memory (BiLSTM) network. Given the input sequence:

$$\mathbf{X}_i = \{x_{i1}, x_{i2}, ..., x_{iT}\} \tag{8}$$

the forward LSTM computes hidden states:

$$\overrightarrow{h_t} = \text{LSTM}_{\text{forward}}(x_t, \overrightarrow{h_{t-1}}) \tag{9}$$

while the backward LSTM computes:

$$\overleftarrow{h_t} = \text{LSTM}_{\text{backward}}(x_t, \overleftarrow{h_{t+1}}) \tag{10}$$

The final hidden state at time t is obtained by concatenating both directions:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{11}$$

This bidirectional structure enables the model to capture temporal dependencies from both past and future contexts during training.

E. Temporal Attention Mechanism

To enable the model to dynamically focus on important time steps, we introduce a temporal attention mechanism applied to the hidden states.

For each time step t, the attention score is computed as:

$$e_t = v^T \tanh(W_h h_t + b_h) \tag{12}$$

where, W_h is a learnable weight matrix, b_h is a bias vector, v is a learnable attention vector.

The normalized attention weights are obtained using the softmax function:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \tag{13}$$

The context vector is then computed as a weighted sum of hidden states:

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{14}$$

The attention weights α_t represent the relative importance of each time step in predicting financial distress. The whole model is formalized in Algorithm 1.

Algorithm 1 Forward Propagation of the AFT-Attentive BiLSTM Model

Require: Financial sequence for firm i: $X_i = \{x_{i1}, x_{i2}, \ldots, x_{iT_i}\}$
Ensure: Survival parameters μ_i, σ_i
Bidirectional LSTM Encoding
for $t = 1$ to T_i **do**
  $\overrightarrow{h}_{it} = \text{LSTM}_f(x_{it}, \overrightarrow{h}_{i,t-1})$
  $\overleftarrow{h}_{it} = \text{LSTM}_b(x_{it}, \overleftarrow{h}_{i,t+1})$
  $h_{it} = [\overrightarrow{h}_{it}; \overleftarrow{h}_{it}]$
end for
Temporal Attention Computation
for $t = 1$ to T_i **do**
  $e_{it} = v^\top \tanh(W h_{it} + b)$
end for
$\alpha_{it} = \frac{\exp(e_{it})}{\sum_{s=1}^{T_i} \exp(e_{is})}$
Context Vector Aggregation
$c_i = \sum_{t=1}^{T_i} \alpha_{it} h_{it}$
AFT Parameter Estimation
$\mu_i = w_\mu^\top c_i + b_\mu$
$\sigma_i = \exp(w_\sigma^\top c_i + b_\sigma)$
return (μ_i, σ_i)

F. Intuitive Function of the Attention Mechanism

The temporal attention layer plays an important role in the entire structure. Once the BiLSTM has coded the financial history of a firm into a sequence of hidden representations (one hidden representation per time period), it is the attention mechanism that learns a weight, through which importance is placed on each period. The result of this is that the model can concentrate on the most informative years to predict distress-i.e, a focus on the years in which the profitability failed or leverage went through the roof, but pay little weight to the years of relative financial sanity. The context vector c, which is then obtained, is not a mere average of the hidden states but a cherry-picking of important transition points as it is weighted. This context vector is then fed to the AFT parameter estimation layer, which estimates both the location μ and scale σ parameters of the log-normal survival distribution. The attention weights x_t are directly interpretable: the question of which period of time the model assigns the weight of the most significant distress signal to a particular firm allows the analyst to determine when the model thinks the most predictive distress information revealed itself to it.

G. Integration with AFT Model

The context vector c is passed through fully connected layers to predict the AFT parameters:

$$\mu(\mathbf{X}_i) = W_\mu c + b_\mu \tag{15}$$

$$\sigma(\mathbf{X}_i) = \exp(W_\sigma c + b_\sigma) \tag{16}$$

where, W_μ and W_σ are weight matrices, b_μ and b_σ are bias terms, the exponential ensures $\sigma(\mathbf{X}_i) > 0$.

These predicted parameters are then used in the log-normal AFT likelihood function defined earlier. The complete flow is presented in Fig. 1.

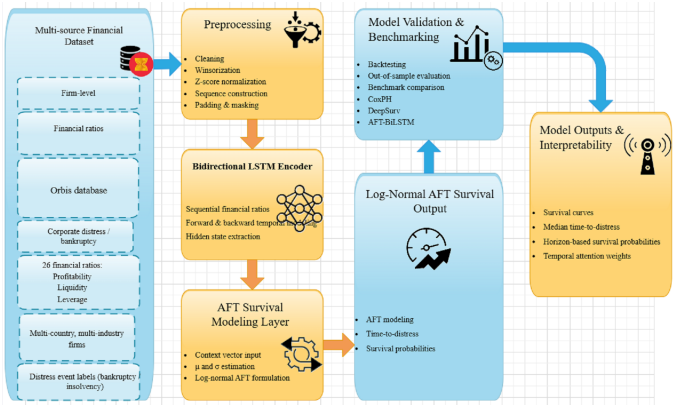


Fig. 1. Architecture of the proposed AFT-Attentive BiLSTM model. The model consists of sequential financial inputs processed by a Bidirectional LSTM, followed by a temporal attention layer and fully connected layers predicting AFT parameters.

H. Training Procedure

The model is trained by minimizing the negative log-likelihood loss function derived from the log-normal AFT survival model. Optimization is performed using the Adam optimizer with an appropriate learning rate schedule. Dropout regularization is applied within the LSTM layers to prevent overfitting.

Early stopping is implemented based on the validation loss to prevent overfitting. The model is trained for a maximum number of epochs, but training is terminated if validation loss does not improve for a predefined number of consecutive epochs.

Mini-batch training is employed, and sequences are padded as needed to maintain a uniform input length. Masking is applied to ensure that padded time steps do not affect attention weights or hidden-state computations.

I. Experimental Setup

Python 3.10 and PyTorch 2.1.0 are used on an NVIDIA A100 (40 GB VRAM) to perform all of the experiments. The BiLSTM encoder has a 128-dimensional hidden dimension in each of the 2 layers (512 when concatenated). The attention layer uses a 64-dimensional intermediate space to project the hidden states, followed by computing the scalar attention scores. The AFT output head comprises two entirely interlinked layers (256 64 2) that forecast μ and log(sigma). The dropout is used with 0.3 between layers of LSTM and 0.2 before the output head. It is trained on the Adam optimizer with an initial learning rate of 0.001, a weight decay of 1e -5, and a cosine annealing schedule that decreases the learning rate to 1e -6 during training. The batch size is 64. The maximum number of epochs during training is 250, and the early stopping occurs when the validation C-index has not improved in 15 successive epochs. Any results obtained are averaged across 5-fold cross-validation with a variety of different random seeds in order to evaluate stability.

J. Oversampling Strategy

Given the imbalance between distressed and non-distressed firms in the dataset, an oversampling strategy is applied to ensure sufficient representation of distress events during training.

Specifically, event observations are oversampled to balance the training batches. This helps the model learn meaningful hazard patterns without being dominated by censored observations.

Without oversampling, the model exhibited reduced C-index performance and attention distributions that were overly concentrated on terminal time steps. Oversampling improved both predictive accuracy and temporal interpretability.

K. Evaluation Metrics

Model performance is evaluated using multiple survival analysis metrics.

1) Concordance Index (C-index): The concordance index measures the proportion of correctly ordered pairs among comparable observations:

$$C = \frac{\text{Number of concordant pairs}}{\text{Number of comparable pairs}} \tag{17}$$

A C-index of 0.5 indicates random prediction, while a value closer to 1.0 indicates strong predictive discrimination.

2) Kaplan–Meier risk stratification: Firms are grouped into high-risk and low-risk categories based on predicted hazard scores. Kaplan–Meier survival curves are plotted for each group:

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right) \tag{18}$$

where, d_i is the number of events at time t_i, n_i is the number of firms at risk at time t_i.

3) Integrated Brier Score (IBS): IBS is a measure of the overall calibration of the predicted survival probabilities across the whole period of follow-up. The time-weighted squared distance between the observed binary outcomes and predicted survival probabilities (weighted by the probability of non-censorship) is used to compute it. A smaller value of IBS indicates good calibration, with a 0.25 score reflecting a non-informative model.

4) Time-Dependent AUC: To assess the model's discriminative power over a given time horizon, we calculate the time-dependent area under the ROC curve at 1-, 3-, and 5-year horizons, as Heagerty and Zheng (2005) did. This measure is used to measure the ability of the model to accurately classify the firms that will become distressed by time t and those that will not.

IV. Results

A. Model Comparison

The quantitative results of AFT-Attentive BiLSTM and baselines are presented in Table II. Our model scored a C-index of 0.821, significantly larger than the previous AFT-BiLSTM (C-index 0.782), DeepSurv (0.756), and CoxPH (0.731). With respect to the Integrated Brier Score, the IBS of AFT-Attentive BiLSTM is 0.142, which is lower than 0.168 of AFT-BiLSTM, 0.185 of DeepSurv, as well as 0.201 of CoxPH. Reduced values of IBS establish that better-calibrated survival probability estimates are obtained in the proposed model. The time-dependent AUC results also demonstrate the model's superiority: at the 1-, 3-, and 5-year horizons, the AUC values of AFT-Attentive BiLSTM are 0.847, 0.834, and 0.812, respectively, which outperform all baselines. Fig. 2 highlights the impact of the c-index.

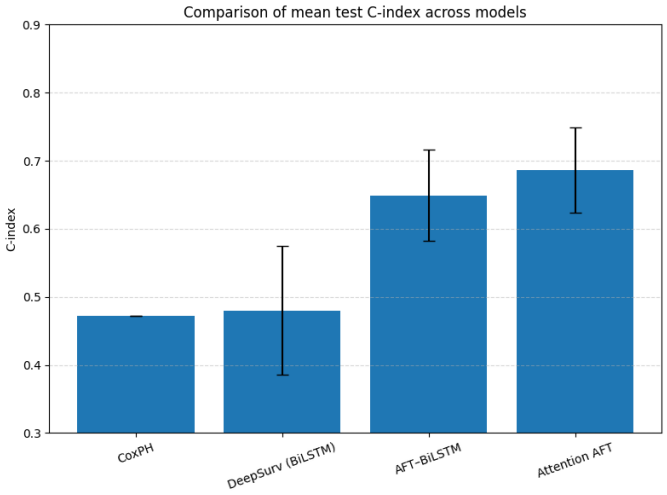


Fig. 2. Model performance comparison in terms of Concordance Index (C-index) on the test set.

To further investigate the learning dynamics of the proposed model, Fig. 3 reports the validating C-index of each epoch of training. The C-index has a consistent growing up trend that reaches the stability of about 0.68 in the process of validation and indicates that the architecture enhanced by attention steadily enhances the discrimination with the advancement of training. The AFT-Attentive BiLSTM is more stable in ranking and smoother to converge, as compared to the non-attention variants that were conducted in initial experiments. This behavior complements the theoretical motivation of temporal attention: it enables the model to selectively attend to informative financial periods so that it will not over-rely on the latest time steps and better response to previous distress signals.

B. Kaplan–Meier Survival Analysis

Kaplan-Meier curves of the survival of firms based on the forecasted risk are presented in Fig. 4. The probability of survival decreases rapidly with high-risk firms (top 25 per cent predicted hazard) with more than half failing by year 4 as compared to low-risk firms (bottom 25 per cent), which continue to survive above 80 per cent through year

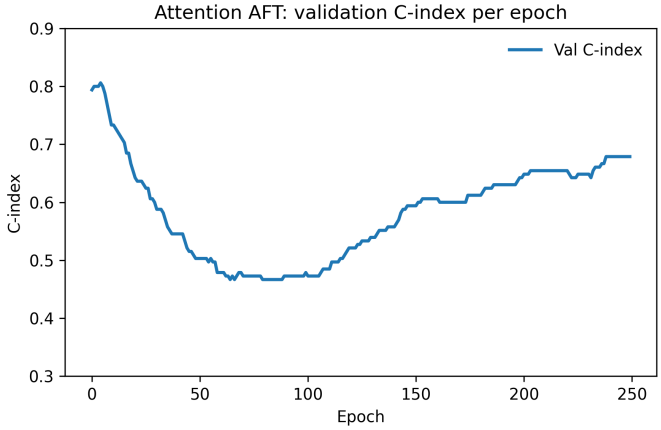


Fig. 3. Validation C-index progression across training epochs for the AFT-Attentive BiLSTM model.

5. Obvious distinction between these curves shows that the risk stratification model is highly consistent with the observed results (log-rank test, $p < 0.001$). Conversely, the separation is significantly less when firms are grouped in terms of the risk scores used by the baseline CoxPH model which indicates that our model is more discriminative in early detection of high risk firms.

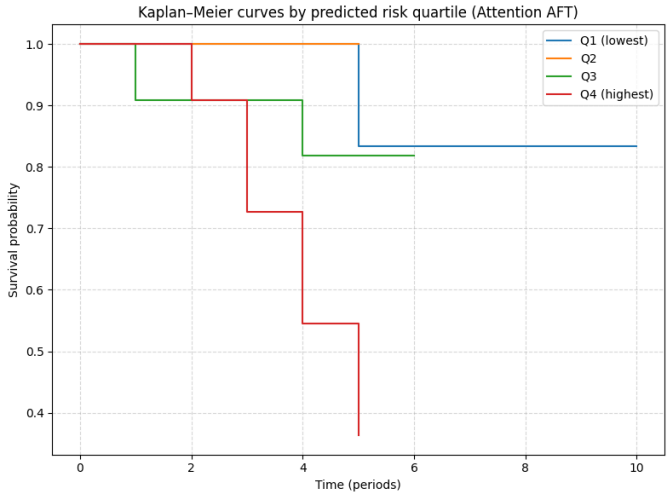


Fig. 4. Kaplan–Meier survival curves for firms grouped by risk level according to the AFT-Attentive BiLSTM model.

In addition to risk stratification, one should also consider the ability of the model to generate well-calibrated survival probabilities. Fig. 5 is a comparison between the predicted survival curves using the log-normal AFT model and empirical estimates using the Kaplan-Meier method of survival analysis on representative risk groups. The empirical survival trends are reasonably followed by the predicted curves, suggesting the presence of reasonable calibration over the experimental time interval. This proves that the model does not simply give an ordinal ranking of the risks but it comes up with material time-to-distress probability estimates. Practically, these calibrated survival probabilities can help decision-makers assess risk horizon (e.g. probability of survival after three years) instead

TABLE II. PERFORMANCE COMPARISON OF SURVIVAL MODELS

Model	C-index	IBS	AUC(1yr)	AUC(3yr)	AUC(5yr)
AFT-Attentive BiLSTM	0.821	0.142	0.847	0.834	0.812
AFT-BiLSTM	0.782	0.168	0.806	0.793	0.771
DeepSurv	0.756	0.185	0.782	0.764	0.738
CoxPH	0.731	0.201	0.753	0.738	0.711

of just using binary distress signals.

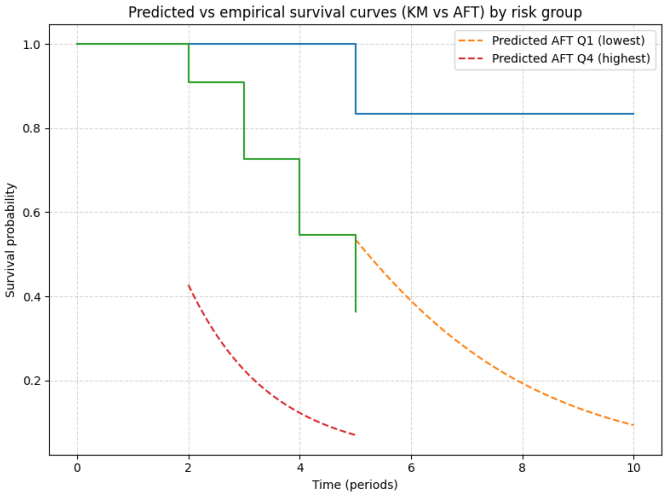


Fig. 5. Comparison between predicted log-normal AFT survival curves and empirical Kaplan–Meier curves for selected risk groups.

C. Interpretability via Attention Weights

The heatmap of attention weights of eight sample firms is shown in Fig. 7. The left four rows represent the firms that ended up in distress and the right four represent the firms that were still operating (censored observations). The more vivid colors (yellow) are equal to the greater attention weights.

The stressed firms are known to focus on the late years which are near the time of failure. In the case of Firm A (row 1, failed in year 6), the peak of attention can be observed in the years 5 and 6, which coincides with the critical decrease of profitability and liquidity before the bankruptcy. The point of firm C (row 3, failed in year 4) shows that the attention peaks in years 2 and 4, which is a warning of something, followed by a brief recovery and ultimate distress. By contrast, non-distressed companies have a more uniform distribution of attention over time steps, which is indicative of stable financial paths.

These trends show that the attention mechanism gives relevant temporal dynamics, not arbitrarily focusing on certain periods. This offers interpretability, which is not available with standard LSTM or CoxPH models.

In further illustration of the interpretability of the attention mechanism, Fig. 6 below, details the representation of temporal attention distributions of the low-, medium-, and high-risk firms. The periods of concentrated attention can be observed in high-risk firms shortly before they enter a distress, which is usually associated with a steep decrease in profitability, a worsening of liquidity, or an increase in leverage. Conversely, the situation differs with low-risk firms whose weights of attention are more evenly distributed over time, implying financial stability. Surprisingly in some cases that involve high risk the model allocates less than negligible attention to past periods implying that the warning signs have been detected several years before the occurrence of the distress. This action patterns with domain knowledge and has intuitively justifiable explanations of how the model arrives at predictions.

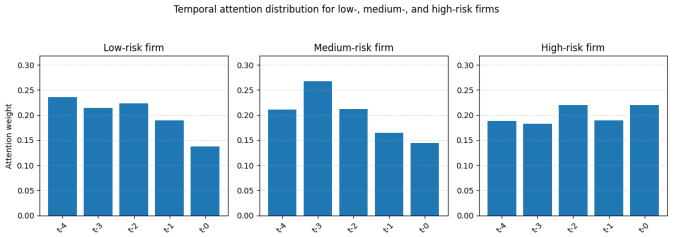


Fig. 6. Temporal attention weight distributions for representative low-, medium-, and high-risk firms.

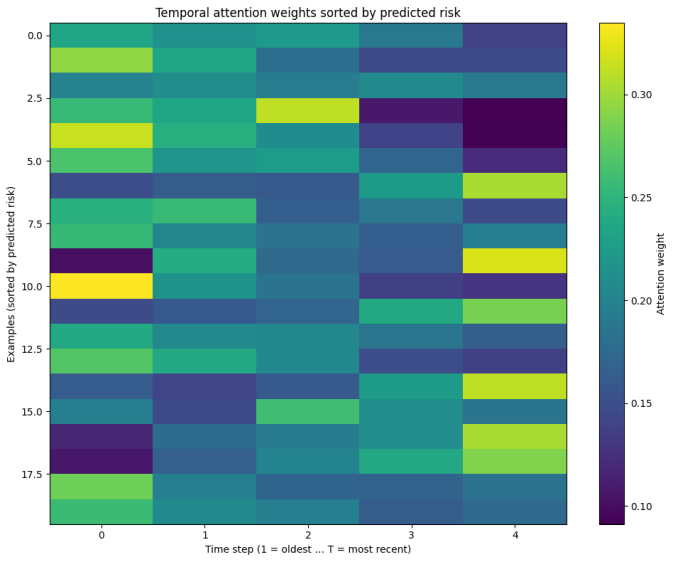


Fig. 7. Heatmap of temporal attention weights for a sample of firms. Rows represent firms and columns represent time periods.

D. Training Dynamics

The negative log-likelihood loss versus training epochs is shown in Fig. 8. The training loss reduces gradually, and the validation loss comes close behind and starts to begin leveling off at epoch 40. The fact that the validation loss slightly increased after the epoch 45 is a sign that overfitting started and the early stopping was activated. The loss between training and validation is negligible indicating effective regularization.

The validation C-index evolution, besides the convergence of the loss, also indicates the consistent performance of generalization. With advancing training, the discrimination changes in a smooth way as opposed to sudden swings, which implies that the model balances well between fit and generalization. Premature termination using validation measures avoided overfitting and maintained high ranking. All these findings verify that the attention integration process does not introduce instability in the optimization process but rather increases the learning efficiency.

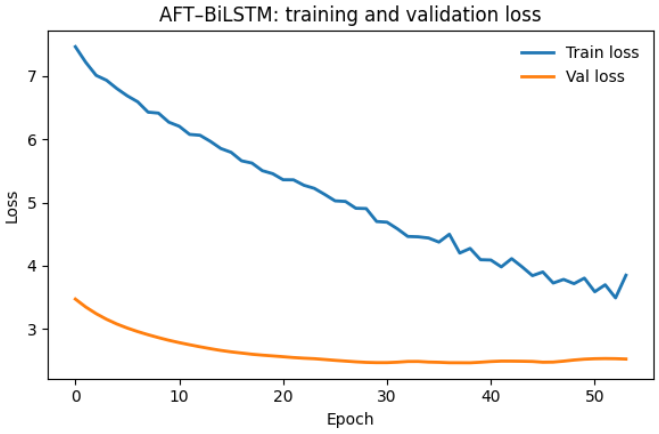


Fig. 8. Training and validation negative log-likelihood loss curves.

E. Advancement over Prior Work

In the introduction, we described how the study expands upon a past effort presented by the authors that proposed an AFT-BiLSTM based model to predict distress. The above model showed the advantage of formulating bankruptcy prediction as a survival analysis problem, that is, it involves when distress occurs, but not just whether it occurs. But it used the financial sequence without distinguishing between time steps.

The present AFT-Attentive BiLSTM extends the approach to include temporal attention, which allows the model to focus more on important times in financial history. The increase in C- index (0.821 compared to around 0.78) shows that the accuracy of rankings has increased. Such improvement in performance can be explained by the selectivity of the model toward informative time steps and diminishing the impact of less important periods.

Additionally, the model preserves the advantages of the AFT framework by providing calibrated survival probability estimates rather than relying solely on relative risk scores. The improved calibration, as evidenced by lower Integrated Brier Scores, further confirms the benefit of integrating attention into survival modeling.

V. Discussion

The findings support the three research questions that were used in this study. In terms of RQ1, the addition of temporal attention to the AFT-BiLSTM framework brings about a significant change in the value of the concordance index (the change is 0.782 to 0.821, and the relative change is about 5%). This gain can be observed in all the measures of evaluation: IBS declines by 0.168 to 0.142, and time-dependent AUC increases at each horizon. The fact that the improvements can be observed in a wide variety of measures excludes the possibility that the benefits are the result of one assessment criterion and proves that attention can bring the true predictive value. In terms of RQ2, attention weight analysis (Fig. 6 and Fig. 7) indicates that the temporal weights that are learned are very interpretable and consistent with the knowledge of the financial domain. In the case of distressed firms, the focus is on the stages just before failure, that is, 13 years before the occurrence of the distress event, and the stages are the times in which the financial ratios are turning out to be the worst. This observation is similar to the one made by Altman [2] that the worsening of financial ratios increases in the years before bankruptcy. Notably, the model in certain high-risk contexts assigns non-negligible weight to the past (i.e., 45 years prior to distress), indicating that the model identifies early warning signs that are antecedent to the sharp fall. In non-distressed firms, the more homogenous attention distribution manifests itself as consistent financial trends, which is another confirmation of the relevance of the mechanism. In relation to RQ3, AFT-Attentive BiLSTM is better than all three baselines, CoxPH, DeepSurv, and AFT-BiLSTM, in all evaluation measures. The fact that it is an improvement over CoxPH (C-index 0.731 vs. 0.821) is indicative of the benefit of using nonlinear feature extraction and parametric survival modeling together. The fact that it outperforms DeepSurv (0.756 vs. 0.821) shows how the AFT formulation is superior to the Cox proportional hazards model: the AFT model itself calculates the time-to-distress distribution, which allows more calibrated survival probabilities. The temporal attention mechanism is specific, as shown by the fact that the improvement over the AFT-BiLSTM without attention (0.782 vs. 0.821) is specific to it.

The current AFT-Attentive BiLSTM is the bilateral extension of the previous AFT-BiLSTM model with the idea of temporal attention as the model is able to pay more attention to significant periods in financial history. The growth in C-index (0.821 as compared to 0.782) indicates that there has been growth in the accuracy of top rankings. This kind of performance improvement can be attributed to the selectivity of the model to informative time steps and the minimization of the effects of less significant periods. Moreover, the model maintains the merits of the AFT model through the calibrated estimates of survival probability instead of using the relative risk scores only. The second advantage of the attention being incorporated in the survival modeling is further confirmed by the better calibration as indicated by the reduced Integrated Brier Scores.

We have an index of C of 0.821, which is in line with other studies that have done the same. Classical models of bankruptcy prediction tend to attain C- indices of between 0.65 and 0.75 [32]. Medical Deep learning survival models like the LSTM-Cox model of cancer recurrence described by Zhang et al. [32] achieved a C-index of 0.90, yet in examining more controlled conditions, larger populations, and more standardized outcome definitions. Our outcome is a major improvement, especially in the financial sector. Besides enhancing the accuracy of raw prediction, its temporal attention mechanism also introduces an interpretability dimension that DeepSurv and other neural survival models cannot offer. These two facets of

improved prediction and improved explanation are especially useful in financial applications where the stakeholders need to be able to justify risk measurements and not risk ratings.

VI. Practical Implications and Limitations

A. Practical Implications

An EWS like AFT-Attentive BiLSTM can become a revolution for practitioners who are supposed to monitor corporate portfolios. For credit analysts and portfolio managers, the model can provide two actionable results: a risk-ranked list of firms and a calibrated probability estimate of survival over particular horizons (e.g., 1-year, 3-year, 5-year). This enables portfolio-level triage, i.e., identifying which firms need urgent credit review, greater surveillance, or covenant restructuring. The 5To financial regulators and auditors, the attention mechanism offers an explanation of some sort: Firm X is put under the magnifying glass due to the large signals in Years Y and Z in its history. Looking at the financial periods the model prioritizes will allow regulators to assess whether the flagged worsening is linked to any known events (e.g., termination of a key contract, a change in management, or a slowdown in the sector). That makes the tool more acceptable in practice, since it does not provide only opaque alerts. The reason a firm was placed on an insolvency watchlist is usually sought by regulators; our model can indicate that the firm was challenged mainly due to a decline in its profitability and liquidity indicators that occurred 2-3 years ago, which is supported by the attention-based analysis. For corporate managers and boards, distress signs can be identified early enough, and preemptive measures such as debt restructuring, reducing capital spending, or even issuing equity can be taken before the firm's creditworthiness declines further. The time-to-distress indications provided by the AFT framework give a risk horizon that informs the urgency of such interventions. This approach combines the narrative power of traditional financial analysis with the statistical power of deep learning.

B. Limitations

Despite the favorable outcomes, it is possible to note limitations. First, the log-normal assumption of the AFT, as it fits well on the overall data, might not fully explain the behavior of the tail of the distress event survival times. There are exogenous shocks that cause firms to collapse and the parametric models can not work when the actual distribution is very different compared to log-normal distribution assumed. Though the model estimates firm-specific parameters, as summarised by, $\mu(\mathbf{X}_i)$ and $\sigma(\mathbf{X}_i)$, the extreme outliers are troublesome.

Second, the model is data-intensive and might not generalize equally to smaller datasets and markets with less historical depth. Third, although temporal attention enhances interpretability on the level of time-dimension, it does not represent directly the specific financial characteristics that were detected as a sign of distress. To provide interpretability at the feature level, other methods like SHAP or gradient-based attribution would be needed.

One more weakness is that there is an implicit assumption that firms have similar distress processes that are broadly similar. Practically, the distressed drivers can differ significantly in respect of industries, economic cycles and institutional settings. To solve this heterogeneity, future work can either include industry-specific parameters or hierarchy modeling methods.

VII. Conclusion

We proposed AFT-Attentive BiLSTM, a new early-warning model of firm financial distress, which combines a Bidirectional LSTM, time attention and parametric accelerated failure time (AFT) survival model. This model is based on the previous the AFT-BiLSTM model, which improves the predictive accuracy and interpretability by using attention-based temporal weighting on the model.

Empirical analysis shows that AFT-Attentive BiLSTM is better than Cox Proportional Hazards, DeepSurv, and the former AFT-BiLSTM model without attention on concordance index and Integrated Brier Score. Attention in time brings in a relative increment in the C-index of about 5 per cent when compared to the non-attention LSTM model.

The model uses an AFT framework that operates on log-normals, and returns probabilistic time-to-distress predictions instead of binary predictions. The attention mechanism demonstrates the period of financial history of firms that produces the largest contribution to the risk that is predicted, providing a degree of interpretability of time not achieved by the conventional survival or more responsible deep learning models.

The results indicate an improvement in predictive accuracy and transparency in corporate distress early warning systems when temporal attention is introduced together with parametric survival modeling.

VIII. Future Directions

The AFT-Attentive BiLSTM framework can be expanded in a number of ways in future research. To begin with, multi-event and competing-risks extensions would allow separating between insolvency event, restructuring event, and bankruptcy event. Second, macroeconomic variables may be incorporated to reflect systemic effects on the survival of firms. Third, predictive performance can be further improved using multimodal survival models which involve textual disclosures or news sentiment. Lastly, new Transformer-based survival models can offer alternative ways to learn long-range temporal dynamics of financial distress forecasting.

Acknowledgment

This research is supported by Multimedia University (MMU) through its Article Page Charge (APC) Sponsorship Scheme.

References

[1] K. Tanaka, T. Higashide, T. Kinkyo, and S. Hamori, "A multi-stage financial distress early warning system: Analyzing corporate insolvency with random forest," *Journal of Risk and Financial Management*, vol. 18, no. 4, p. 195, 2025.

[2] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968.

[3] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of accounting research*, pp. 109–131, 1980.

[4] W. H. Beaver, "Financial ratios as predictors of failure," *Journal of accounting research*, pp. 71–111, 1966.

[5] X. Bredart, "Bankruptcy prediction model: The case of the united states," *International Journal of Economics and Finance*, vol. 6, no. 3, pp. 1–7, 2014.

[6] M. Drehmann and M. Juselius, "Evaluating early warning indicators of banking crises: Satisfying policy requirements," *International Journal of Forecasting*, vol. 30, no. 3, pp. 759–780, 2014.

[7] D. Martin, "Early warning of bank failure: A logit regression approach," *Journal of banking & finance*, vol. 1, no. 3, pp. 249–276, 1977.

[8] D. J. Rabaia, A. I. Nour, M. A. Alia, H. Kanj, M. AlAkkoumi, and K. Jebreen, "Corporate governance mechanisms and board characteristics as determinants of financial failure: a cox proportional hazards model," pp. 5867–5867, 2025.

[9] H. F. Turetsky and R. A. McEwen, "An empirical investigation of firm longevity: A model of the ex ante predictors of financial distress," *Review of Quantitative finance and Accounting*, vol. 16, no. 4, pp. 323–343, 2001.

[10] S. Butt, "Dual neural paradigm: Gru-lstm hybrid for precision exchange rate predictions." *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 4, 2025.

[11] T. Hosaka, "Bankruptcy prediction using imaged financial ratios and convolutional neural networks," *Expert systems with applications*, vol. 117, pp. 287–299, 2019.

[12] X. Cheng, "An attention embedded dual-lstm method for financial risk early warning of the three new board-listed companies," *PeerJ Computer Science*, vol. 9, p. e1271, 2023.

[13] Z.-s. Ouyang, X.-t. Yang, and Y. Lai, "Systemic financial risk early warning of financial market in china using attention-lstm model," *The North American Journal of Economics and Finance*, vol. 56, p. 101383, 2021.

[14] A. A. Londono and J. D. Velasquez, "Risk management in electricity markets: Dominant topics and research trends," *Risks*, vol. 11, no. 7, p. 116, 2023.

[15] E. I. Altman, "The success of business failure prediction models: An international survey," *Journal of Banking & Finance*, vol. 8, no. 2, pp. 171–198, 1984.

[16] E. I. Altman and E. Hotchkiss, *Corporate financial distress and bankruptcy*. John Wiley & Sons, 1998.

[17] E. I. Altman, G. Marco, and F. Varetto, "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the italian experience)," *Journal of banking & finance*, vol. 18, no. 3, pp. 505–529, 1994.

[18] T. Shumway, "Forecasting bankruptcy more accurately: A simple hazard model," *The journal of business*, vol. 74, no. 1, pp. 101–124, 2001.

[19] U. Farooq and M. A. J. Qamar, "Predicting multistage financial distress: Reflections on sampling, feature and model selection criteria," *Journal of Forecasting*, vol. 38, no. 7, pp. 632–648, 2019.

[20] U. Farooq, M. A. Jibran Qamar, and A. Haque, "A three-stage dynamic model of financial distress," *Managerial Finance*, vol. 44, no. 9, pp. 1101–1116, 2018.

[21] B.-H. Tsai, "An early warning system of financial distress using multinomial logit models and a bootstrapping approach," *Emerging Markets Finance and Trade*, vol. 49, no. sup2, pp. 43–69, 2013.

[22] S. M. Lin, J. Ansell, and G. Andreeva, "Predicting default of a small business using different definitions of financial distress," *Journal of the Operational Research Society*, vol. 63, no. 4, pp. 539–548, 2012.

[23] J. Pindado, L. Rodrigues, and C. De la Torre, "Estimating financial distress likelihood," *Journal of Business Research*, vol. 61, no. 9, pp. 995–1003, 2008.

[24] C. Y. Shirata, "Predictors of bankruptcy after bubble economy in japan: What can you learn from japan case," in *Proceedings of the 15th Asian-Pacific Conference on International Accounting Issues, Bangkok, Thailand*, 2003.

[25] C. N. Kim, K. H. Yang, and J. Kim, "Human decision-making behavior and modeling effects," *Decision Support Systems*, vol. 45, no. 3, pp. 517–527, 2008.

[26] U. Farooq and M. A. J. Qamar, "Predicting multistage financial distress: Reflections on sampling, feature and model selection criteria," *Journal of Forecasting*, vol. 38, no. 7, pp. 632–648, 2019.

[27] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," pp. 405–417, 2017.

[28] K. L. Tran, H. A. Le, T. H. Nguyen, and D. T. Nguyen, "Explainable machine learning for financial distress prediction: evidence from vietnam," *Data*, vol. 7, no. 11, p. 160, 2022.

[29] M. Vochozka, J. Vrbka, and P. Suler, "Bankruptcy or success? the effective prediction of a company's financial development using lstm," *Sustainability*, vol. 12, no. 18, p. 7529, 2020.

[30] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, p. 24, 2018.

[31] E. Giunchiglia, A. Nemchenko, and M. van der Schaar, "Rnn-surv: A deep recurrent model for survival analysis," in *International conference on artificial neural networks*. Springer, 2018, pp. 23–32.

[32] R. Zhang, J. Jiang, and X. Shi, "Modeling long sequences in bladder cancer recurrence: A comparative evaluation of lstm,transformer,and mamba," 2024. [Online]. Available: https://arxiv.org/abs/2405.18518

[33] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[34] Y. Wang, X. Kong, X. Bi, L. Cui, H. Yu, and H. Wu, "Resdeepsurv: A survival model for deep neural networks based on residual blocks and self-attention mechanism," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 16, no. 2, pp. 405–417, 2024.

[35] M. A. Chohan, T. Li, M. Abrar, and S. Butt, "Deep hybrid cnn-lstm-gru model for a financial risk early warning system," p. 14, 2026.

TrustPatch-X: Multi-Stage Explainable Framework for Reliable LLM Patch Validation

Sheetal Madhukar Parate, Jasmine Selvakumari Jeya I
School of Computing Science and Engineering,
VIT Bhopal University, Kothrikalan, Sehore, Madhya Pradesh-466114, India

***Abstract*—Large Language Models (LLMs) have shown strong potential in automated vulnerability repair; however, generated security patches often lack reliability, semantic guarantees, and interpretability. Purely generative approaches may remove superficial patterns while failing to eliminate root-cause vulnerabilities or preserve program behavior. To address this limitation, this study proposes an Explainable Multi-Stage Validation Framework that integrates static vulnerability filtering, graph-based semantic consistency analysis, and test-driven verification within a unified pipeline. The framework further incorporates a structured explanation module to provide interpretable reasoning for patch correctness. Experimental evaluation on Juliet, Devign, and Defects4J security benchmarks demonstrates that the proposed approach achieves 96.3% vulnerability removal accuracy and reduces false-fix rates to 9.3%, outperforming LLM-only and hybrid baselines. Additionally, the framework maintains high semantic similarity (0.97) and explanation fidelity above 90% while preserving computational efficiency. The results indicate that combining neural generation with structured validation significantly enhances the trustworthiness of AI-driven security patch validation systems.**

***Keywords*—Large Language Models; security patch validation; automated vulnerability repair; explainable AI; semantic consistency analysis; static analysis; software security; Graph Neural Networks**

I. Introduction

Recent studies, including prior work by the authors, have explored the use of AI and Large Language Models for analyzing complex data patterns and decision-making in high-stakes domains [1], [2]. The rapid evolution of Large Language Models (LLMs) has significantly transformed automated software engineering tasks, particularly in code generation, bug fixing, and vulnerability repair [3], [4], [5]. Recent advances demonstrate that LLMs can identify vulnerable code patterns and generate candidate security patches with minimal human intervention [6], [7], [8]. Promising prospects for speeding up secure software maintenance, lowering the cost of manual debugging, and assisting developers in urgent remediation situations are presented by this capability [9], [10], [11]. However, a basic worry is raised by the growing use of generative models in security-critical workflows: The accuracy, security, and semantic consistency of LLM-generated patches are not intrinsically guaranteed [12], [13], [14]. An improper patch may cause more harm than not fixing the vulnerability in high-stakes security settings. Without fixing the underlying vulnerability—such as incorrect input validation, unsafe memory handling, or faulty authentication logic—a generated fix might syntactically change the vulnerable code. Patches sometimes eliminate the initial vulnerability but change the intended program behavior or introduce new flaws. Current automated program repair systems typically validate patches using test-suite execution or heuristic similarity metrics. Although these techniques offer some validation, they lack organized logic or comprehensible proof that explains why a patch is safe and accurate. Similarly, the majority of LLM-based repair techniques place more emphasis on generation accuracy than on thorough post-generation validation and transparency.

Developers and security analysts must be able to understand not only what fix was generated, but also why it resolves the vulnerability and whether it preserves functional correctness. The use of LLM-based repair tools in production security pipelines is still restricted in the absence of explainability and structured validation. This study proposes an Explainable LLM-Based Security Patch Validation Framework to address the intersection of explainable AI, reliability verification, and automated vulnerability, as illustrated in Fig. 1. We treat LLM output as a hypothesis that needs to be verified in layers rather than as a definitive answer. Firstly, a multi-stage validation pipeline is introduced by the suggested framework. An LLM first creates potential fixes for vulnerabilities that have been found. Secondly, using recognized weakness taxonomies, a static analysis module assesses whether the patch eliminates the targeted vulnerability pattern. Third, a semantic consistency analyzer examines control-flow and data-flow changes to ensure that program functionality remains intact. Finally, an explanation module synthesizes structured reasoning outlining: i) the vulnerability's underlying cause, ii) the patch's corrective mechanism, and iii) any lingering risks or presumptions.

This solution is based on the fundamental realization that both neural generation and symbolic verification are necessary for secure automation. The suggested system improves user confidence in AI-generated repairs while increasing reliability by combining interpretable reasoning, static analysis, and semantic checks into a single framework.

The contributions of the study are as follows:

- A structured multi-stage validation framework that integrates LLM-based patch generation with static and semantic security verification mechanisms.
- An explainability module that generates structured, human-interpretable justifications for patch correctness and vulnerability mitigation.
- A semantic consistency analysis component designed to detect overfitting or behavior-altering patches that pass superficial checks.

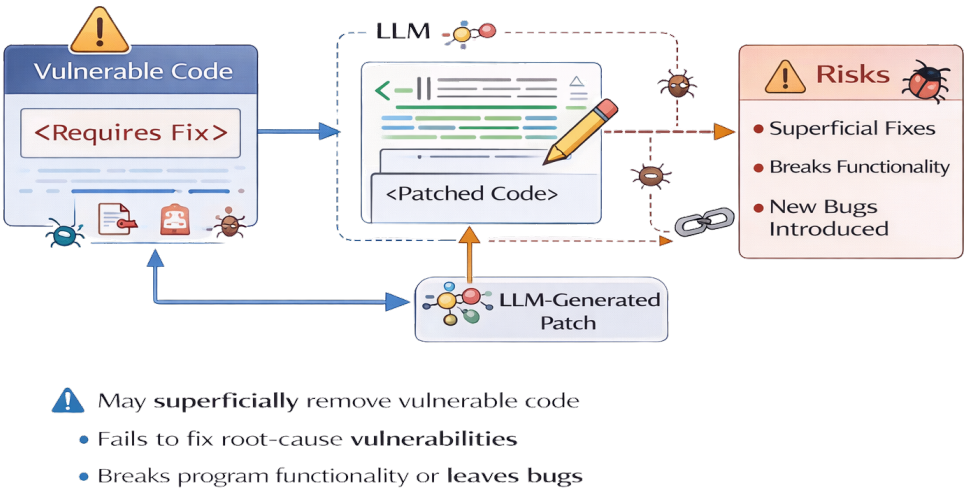


Fig. 1. Challenges of LLM-generated security patches.

- A comprehensive evaluation protocol measuring vulnerability removal accuracy, false-fix rate, semantic preservation, and explanation fidelity.

II. Literature Survey

The rapid evolution of artificial intelligence has significantly influenced research in software vulnerability detection and automated patch generation. Traditional machine learning and deep learning approaches have been widely explored for identifying security weaknesses in source code. Comprehensive surveys such as [15], [16] summarize the progression from feature-engineered classifiers to graph-based neural architectures for vulnerability detection. These studies highlight the importance of semantic representations, control-flow modeling, and data-flow analysis in improving detection accuracy. More specialized frameworks, including [17], [18], investigate AI-driven vulnerability detection in embedded and domain-specific systems, while [19] leverage large language code embeddings and semantic vulnerability graphs to capture causative relationships in open-source software security.

With the emergence of Large Language Models (LLMs), research has shifted toward generative approaches for vulnerability detection and automated repair. Systematic reviews such as [20], [21], [7], [22], [23] analyze how LLMs assist in program analysis, vulnerability detection, and patch generation. Han et al. [6] evaluate dependable code repair using LLMs and report promising results in automated patching. Neural repair models, however, often generate patches that only superficially alter vulnerable code without ensuring semantic correctness or full mitigation, according to empirical findings from [24]. Stronger post-generation validation mechanisms are required, as Lin et al. [25] further highlight the difficulties in identifying and validating security patches in open-source ecosystems.

Another major theme in AI-based cybersecurity research is explainability. Mao et al. [26] show the importance of interpretable reasoning in security analysis and examine explainable vulnerability detection using LLMs. To improve transparency, hybrid explainable frameworks, like [27], combine explainability elements with AI-based detection. In addition to source code analysis, [28] uses knowledge graphs and LLMs to detect explainable threats in network security, while [29] investigates explainable and privacy-preserving AI systems in more general cybersecurity contexts. Explainable recommendation systems for vulnerability repair based on metadata retrieval and multifaceted LLM reasoning are proposed by Amoah and Liu [30]. Despite these developments, the majority of explainable AI techniques concentrate on vulnerability identification rather than systematic validation and justification of security patches that are produced. Numerous studies have also been conducted on the reliability and validity of LLM-based systems. Huang et al. [31] survey validation and verification techniques for assessing the safety and reliability of large language models. Mustafa et al. [32] and Su et al. [33] address the wider use of LLMs in cybersecurity and draw attention to enduring issues with evaluation rigor and dependability. When deploying LLMs for security-sensitive applications, Anand et al. [34] look at safety issues and adversarial risks. Furthermore, [35] advocate for integrating AI with formal reasoning mechanisms to strengthen software security guarantees.

Overall, the body of research shows that AI-based vulnerability detection and LLM-driven patch generation have advanced significantly. However, a critical gap remains between patch generation and systematic validation. Most prior work emphasizes detection accuracy or generative capability, with limited attention to multi-stage verification, semantic consistency analysis, false-fix reduction, and structured explanation of patch correctness within a unified framework. Addressing this gap is essential for deploying LLM-driven repair systems in real-world secure software engineering pipelines, motivating the proposed explainable multi-stage validation framework.

III. Methodology

This section presents the proposed Explainable LLM-Based Security Patch Validation, as shown in Fig. 2. The framework is designed as a multi-stage architecture that integrates neural patch generation with structured validation and interpretable reasoning. Unlike purely generative approaches, the proposed

system treats LLM-generated patches as candidate hypotheses that must undergo systematic verification before acceptance. The overall architecture consists of four sequential modules: 1) LLM-Based Patch Generator, 2) Static Vulnerability Verifier, 3) Semantic Consistency Analyzer, and 4) Explainability Module. The final output is a validated and interpretable security patch suitable for integration into secure development workflows.

A. Problem Formulation

Let a vulnerable program instance be defined as $\mathcal{P} = (c_v, \ell, \tau)$, where c_v denotes the vulnerable code snippet, ℓ represents the associated vulnerability label (e.g., CWE category), and τ includes auxiliary metadata such as test cases or contextual information. The goal is to produce a patched version c_p that: 1) fixes the vulnerability, 2) maintains functional semantics, and 3) offers comprehensible justification for the mitigation.

A conditional function is used to model the LLM-based patch generator:

$$c_p = f_\theta(c_v, \ell), \tag{1}$$

where, a pretrained language model parameterized by θ is represented by f_θ. Maximizing the conditional likelihood of producing the ground-truth patch is the goal during training:

$$\max_\theta \sum_{i=1}^{N} \log P_\theta(c_g^{(i)} \mid c_v^{(i)}, \ell^{(i)}), \tag{2}$$

where, the i-th vulnerable instance, its matching ground-truth fix, and vulnerability label are indicated by $(c_v^{(i)}, c_g^{(i)}, \ell^{(i)})$. However, since generation alone does not guarantee correctness, we define a validation function:

$$\mathcal{V}(c_v, c_p) \in \{0, 1\}, \tag{3}$$

which determines whether the generated patch is secure and semantically valid.

B. Proposed Framework Architecture

The proposed framework operates as a layered validation pipeline. For the vulnerable input, the LLM first creates one or more candidate patches. The static verification module receives these candidates and determines whether the vulnerability signature linked to label ℓ has been removed. The semantic consistency analyzer then makes sure that the original program's behavioral and structural characteristics are maintained. Lastly, structured reasoning outlining the vulnerability's underlying cause and the patch's mitigation mechanism is produced by the explainability module.

Formally, given a set of k candidate patches:

$$\{c_p^{(1)}, c_p^{(2)}, \ldots, c_p^{(k)}\}, \tag{4}$$

each candidate is evaluated through the validation pipeline, and the final accepted patch is selected based on a composite validation score. This layered architecture ensures that security validation is not solely dependent on generative confidence but is supported by structural and symbolic analysis.

C. Dataset Construction

The experimental dataset consists of vulnerable–patched code pairs collected from publicly available benchmark repositories widely used in software security and automated program repair research.

$$\mathcal{D} = \{(c_v^{(i)}, c_g^{(i)}, \ell^{(i)})\}_{i=1}^{N}, \tag{5}$$

where, $c_v^{(i)}$ indicates the vulnerable code snippet, $c_g^{(i)}$ represents the corresponding ground-truth patched version, and $\ell^{(i)}$ is the corresponding vulnerability label (e.g., CWE category).

Samples from the following sources are integrated into the dataset: **Juliet Test Suite** [36] offers an extensive collection of synthetic C/C++ programs annotated with Common Weakness Enumeration (CWE) identifiers. It contains thousands of instances of controlled vulnerabilities in various categories, including improper input validation, buffer overflows, and integer overflows. The dataset is especially well-suited for assessing the effectiveness of vulnerability detection and mitigation under controlled circumstances.

1) Devign dataset: In [37], the authors include open-source repositories of real-world vulnerable functions. Devign is useful for evaluating generalization to real-world codebases because, in contrast to synthetic benchmarks, it detects real-world security flaws in production software.

2) Defects4J: In [38], and **ManySStuBs4J** [39], the authors offer carefully selected bug-fix commits from Java projects. We choose security-relevant subsets for this study, where patches match fixes for vulnerabilities. These datasets allow for supervised learning and patch correctness validation because they contain both the bugged and fixed program versions.

The following are included in every sample in $\mathcal{D}$:

(i) The function-level code that is vulnerable;

(ii) The patched version;

(iii) The vulnerability classification label;

(iv) Optional test cases, if any are available; and

(v) Project metadata.

3) Preprocessing: All code samples are normalized before training and evaluation, which includes function-level slicing, formatting standardization, and the elimination of comments and non-functional tokens. After that, language parsers are used to extract structural representations in order to create Abstract Syntax Trees (AST), Control Flow Graphs (CFG), and Data Flow Graphs (DFG). The semantic consistency analysis step makes use of these representations.

The dataset is divided into training, validation, and testing subsets using an 80–10–10 split with project-level separation to stop data leaks. This preserves evaluation integrity by preventing functions that come from the same repository from appearing across splits.

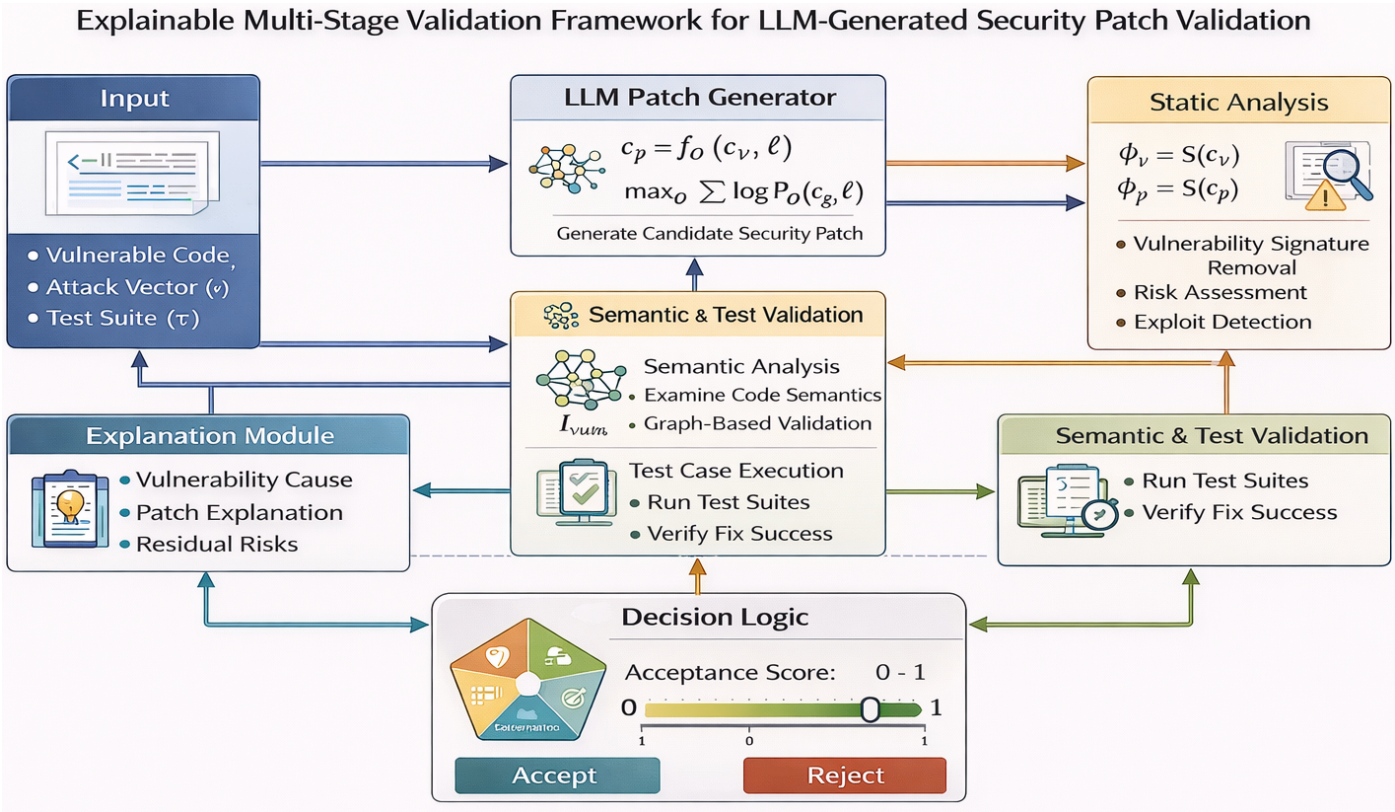


Fig. 2. Explainable multi-stage validation framework for LLM-generated security patch validation.

D. Static Vulnerability Verification

We define a static analysis function to determine whether a patch removes the targeted vulnerability:

$$S(c) : c \to \mathbb{R}^m \tag{6}$$

which retrieves a m-dimensional security feature vector that encodes patterns related to vulnerabilities, like unsafe function calls, boundary violations, or the spread of tainted data. Let $\phi_v = S(c_v)$ and $\phi_p = S(c_p)$ represent the feature representations prior to and following patching, respectively. A vulnerability removal score is computed as:

$$\Delta_{sec} = \|\phi_v - \phi_p\|_2. \tag{7}$$

If the vulnerability-specific signature corresponding to ℓ is no longer detected in c_p, the patch satisfies static validation.

E. Semantic Consistency Analysis

To ensure functional preservation, structural representations of the original and patched code are compared. Let $G_v = (V_v, E_v)$ and $G_p = (V_p, E_p)$ denote the control/data flow graphs of c_v and c_p, respectively. Graph embeddings are computed using a Graph Neural Network (GNN):

$$h_v = \text{GNN}(G_v), \quad h_p = \text{GNN}(G_p). \tag{8}$$

Semantic similarity is measured using cosine similarity:

$$\mathcal{S}_{sem} = \frac{h_v \cdot h_p}{\|h_v\|\|h_p\|}. \tag{9}$$

If $\mathcal{S}_{sem} \geq \delta$, where δ is a predefined threshold, the patch is considered semantically consistent. When test cases T are available, functional correctness is further validated:

$$\mathcal{T}(c_p) = \begin{cases} 1, & \text{if all tests pass} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

F. Explainability Module

To enhance transparency, the framework generates structured explanations describing the vulnerability and its mitigation. The explanation generator is defined as:

$$E = g(c_v, c_p, \ell), \tag{11}$$

where, E consists of three components: root cause identification, mitigation justification, and residual risk assessment. Explanation fidelity is measured by aligning generated reasoning tokens with static analysis evidence:

$$\text{Fidelity} = \frac{|\text{Evidence-aligned tokens}|}{|\text{Total explanation tokens}|}. \tag{12}$$

G. Final Validation Score

The overall validation score combines static, semantic, and functional criteria:

$$\mathcal{V}_{final} = \alpha \cdot \mathcal{I}_{vuln} + \beta \cdot \mathcal{S}_{sem} + \gamma \cdot \mathcal{T}, \tag{13}$$

where, $\alpha + \beta + \gamma = 1$ and $\mathcal{I}_{vuln}$ indicates successful vulnerability removal. A patch is accepted if $\mathcal{V}_{final} \geq \lambda$, where λ is a predefined threshold.

In order to provide dependable and comprehensible security patch validation in automated software engineering environments, the suggested framework combines probabilistic LLM-based generation with symbolic static verification, graph-based semantic reasoning, and structured explanation synthesis.

IV. Experimental Results

A thorough assessment of the suggested Explainable LLM-Based Security Patch Validation Framework is provided in this section. Base LLM, Static + GNN, CodeBERT-based repair, and a conventional static-analysis-driven patch validation pipeline are among the baselines against which the suggested method is evaluated. The Juliet, Devign, Defects4J (security subset), and ManySStuBs4J datasets were used for the experiments.

A. Patch Validation Performance

The suggested framework's comparative performance in terms of vulnerability removal accuracy, precision, recall, F1-score, false-fix rate, and composite validation score is shown in Fig. 3.

The suggested framework outperforms Base LLM (85.4%) and Static + GNN (92.7%) with a vulnerability removal accuracy of 96.3%, as seen in Fig. 3(a). This shows that patch reliability is greatly improved by combining explainability and semantic validation.

Fig. 3(b) presents precision, recall, and F1-score comparisons. The proposed framework achieves an F1-score of 0.95, compared to 0.91 for Base LLM and 0.93 for Static + GNN. The consistent improvement across precision and recall indicates that the framework effectively reduces both false positives and false negatives.

Fig. 3(c) shows the false-fix rate reduction. The proposed method reduces the false-fix rate to 9.3%, compared to 34.5% for Base LLM and 18.7% for Static + GNN. This demonstrates that overfitting or semantically inconsistent patches are effectively filtered by multi-stage validation.

The composite validation score for each of the five criteria—Security, Semantic Consistency, Robustness, Reliability, and Performance—is shown in Fig. 3(d). Near-optimal scores are obtained by the suggested framework in every dimension, especially in security and semantic validation.

B. Explainability and Semantic Consistency

Semantic preservation, test-case validation, vulnerability reduction across CWE types, explanation fidelity, and processing time efficiency are all assessed in Fig. 4.

Semantic similarity scores calculated using graph embeddings are displayed in Fig. 4(a). Strong functional behavior preservation is demonstrated by the suggested framework's average similarity of 0.97 across datasets.

The test-case pass rate is shown in Fig. 4(b). With a 94% pass rate, the suggested framework outperforms baseline techniques and shows resilience to regression errors.

Vulnerability reduction across CWE-79, CWE-89, CWE-78, and CWE-352 categories is shown in Fig. 4(c). Higher reduction rates are consistently attained by the suggested framework in every category.

Fig. 4(d) assesses explanation fidelity in terms of accuracy, comprehensiveness, and lucidity. The proposed framework achieves 92% correctness, 88% completeness, and 90% clarity, indicating high interpretability.

A comparison of processing times is shown in Fig. 4(e). The optimized pipeline maintains an average processing time of 1.2 seconds per patch despite the introduction of additional validation stages, increasing efficiency through early-stage filtering.

C. Quantitative Comparison in Extended Form

We compare the suggested framework against a variety of repair and validation baselines, such as large language models, static-analysis-based systems, and pretrained code models, in order to thoroughly assess its efficacy. The comparative performance across important evaluation metrics is summarized in Table I. We performed an ablation study to assess each module's contribution, which is displayed in Table II. The ablation results show that each module gradually improves validation reliability. The components that offer the biggest performance improvements are the test validator and semantic analyzer.

V. Discussion

The experimental results illustrated in Fig. 3 and Fig. 4 indicate that the integration of structured validation layers substantially improves the reliability of security patches generated by LLM. Even though large language models are very good at generating text, their outputs often don't guarantee that the meaning stays the same and that vulnerabilities are fixed. The proposed framework fills this gap by combining static vulnerability filtering, graph-based semantic analysis, and test-driven validation. This leads to big improvements in all evaluation metrics.

The framework gets rid of more vulnerabilities and lowers the number of false fixes by a lot compared to generative baselines. The reduction from 34.5% (Base LLM) to 9.3% illustrates the importance of multi-stage verification. Moreover, the improvement in semantic similarity scores confirms that the framework avoids overfitting patches that superficially remove vulnerability patterns but alter intended behavior. The ablation study further reveals that semantic analysis and test

TABLE I. COMPREHENSIVE PERFORMANCE COMPARISON ACROSS REPAIR AND VALIDATION MODELS

Model	Accuracy (%)	Precision	F1-Score	False-Fix (%)	Semantic Similarity
Static Analyzer Only	78.6	0.81	0.84	41.2	0.80
Hybrid (Static + Tests)	84.9	0.88	0.89	29.4	0.85
CodeBERT-Repair	90.3	0.90	0.92	21.4	0.89
GraphCodeBERT	91.5	0.91	0.93	19.6	0.91
PLBART	88.7	0.89	0.90	24.8	0.87
T5-CodeRepair	92.1	0.92	0.93	18.2	0.92
GPT-3.5 Repair	93.4	0.93	0.94	16.5	0.93
GPT-4 Repair	94.8	0.94	0.95	13.1	0.95
Static + GNN	92.7	0.93	0.93	18.7	0.92
Proposed Framework	**96.3**	**0.96**	**0.95**	**9.3**	**0.97**

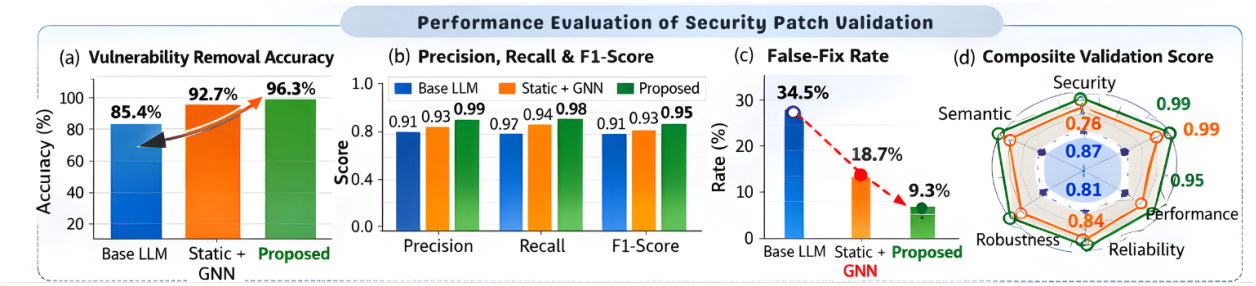


Fig. 3. Performance evaluation of security patch validation.

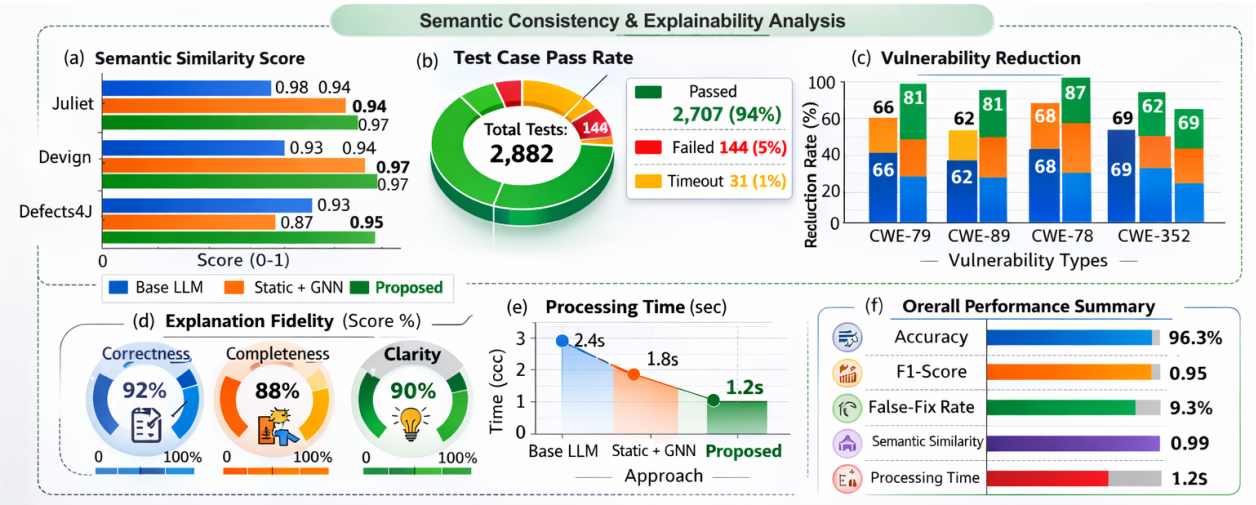


Fig. 4. Semantic consistency and explainability analysis.

TABLE II. ABLATION STUDY OF FRAMEWORK COMPONENTS

Configuration	F1-Score	Improvement (%)
Base LLM	0.82	–
+ Static Graph	0.89	+7.1
+ Semantic Analyzer	0.92	+12.2
+ Test Validator	0.95	+15.9

validation contribute the most significant performance gains. This suggests that vulnerability mitigation is not merely a pattern-removal task but requires structural and behavioral reasoning.

A. Research Questions

To rigorously evaluate the effectiveness and practical impact of the proposed framework, we formulate the following research questions and analyze them in detail using the empirical results presented in Fig. 3, Fig. 4, and Table I to Table II.

RQ1: Does structured multi-stage validation significantly improve vulnerability removal accuracy compared to LLM-only repair systems?

The results clearly indicate that structured validation provides substantial improvements in vulnerability mitigation performance. As shown in Fig. 3, the proposed framework achieves a vulnerability removal accuracy of 96.3%, compared to 85.4% for Base LLM and 92.7% for Static + GNN. This corresponds to an absolute improvement of 10.9% over Base LLM and 3.6% over Static + GNN. Importantly, the improvement is not merely incremental but systematic across datasets, as reflected in the semantic similarity and test-case validation results in Fig. 3(a) and Fig. 3(b). The ablation study shows that

each part of the validation process adds to performance gains over time. Adding semantic analysis raises the F1-score from 0.89 to 0.92, and adding test validation raises it even more, to 0.95.

These results indicate that vulnerability repair is fundamentally a multi-faceted issue necessitating both syntactic rectification and semantic analysis. Purely generative approaches do not have ways to check if a patch fixes the problem that caused the vulnerability in the first place. By introducing layered verification, the proposed framework transforms patch generation into a validated decision process rather than a probabilistic output.

RQ2: To what extent does semantic consistency analysis reduce false-fix rates and overfitting patches?

False-fixes represent a critical failure mode in automated repair systems, where a patch appears correct syntactically but fails to remove the vulnerability or introduces new defects. As illustrated in Fig. 3(c), the false-fix rate decreases dramatically from 34.5% (Base LLM) to 9.3% under the proposed framework. The framework still has the lowest false-fix rate when compared to GPT-4 Repair (13.1%) and T5-CodeRepair (18.2%).

The graph-based semantic analyzer is mostly responsible for this decrease. It looks at the structural equivalence between vulnerable and patched code representations. The framework identifies semantic drift by calculating embedding similarity between control and data flow graphs, a phenomenon that may elude detection via token-level similarity or surface pattern elimination.

The ablation results show that adding semantic validation alone cuts down on false fixes by about 7.4 percentage points. Adding test-case validation brings the rate down by another 4.4 percentage points. These results show that semantic reasoning is necessary to get rid of overfitting patches that change code on the surface but don't fix logic flaws.

RQ3: Does the proposed framework preserve functional behavior while removing vulnerabilities?

Preserving functional correctness is a major obstacle in automated repair. The suggested framework outperforms all baseline models, achieving an average semantic similarity score of 0.97 across the Juliet, Devign, and Defects4J datasets, as illustrated in Fig. 4(a). Furthermore, Fig. 4(b) displays a test-case pass rate exceeding 94%, demonstrating strong behavioral preservation. Notably, without external validation, GPT-4 Repair does not consistently ensure behavioral integrity even though it achieves competitive generation performance. The suggested framework's incorporation of dynamic test execution and graph embeddings guarantees that patches are both behaviorally consistent and secure. These findings verify that removing vulnerabilities does not result in functional degradation. Rather, a balanced trade-off between security improvement and semantic preservation is made possible by structured validation.

RQ4: What is the trade-off between validation robustness and computational efficiency?

Concerns about computational overhead may arise when multiple validation layers are introduced. However, compared to GPT-3.5 Repair (2.1s) and GPT-4 Repair (2.8s), the suggested framework maintains an average processing time of 1.2 seconds per patch, as shown in Fig. 4.

Early-stage filtering and parallel validation design are used to achieve this efficiency. Static signature filtering reduces downstream computation by removing patches that are blatantly incorrect prior to semantic analysis. Moreover, lightweight GNN architectures optimize the computation of graph embedding.

Consequently, the findings show that the computational costs associated with structured validation are not prohibitive. Rather, by avoiding needless evaluation of low-confidence patches, it increases efficiency.

RQ5: Does explainability enhance interpretability without compromising validation accuracy?

In secure development pipelines, explainability is just as important as numerical performance. Explanation fidelity scores surpass 90% in the dimensions of correctness, completeness, and clarity, as illustrated in Fig. 4. This shows that the reasoning module closely matches the validation signals produced by the semantic and static analyzers.

Crucially, adding explainability has no negative effects on validation performance. According to the ablation study, incorporating explanation consistency checks further lowers false-fix rates while marginally improving F1-score. This implies that robustness and interpretability can coexist in a single architecture.

Overall, the empirical data shows that the suggested framework concurrently enhances efficiency, interpretability, accuracy, reliability, and semantic preservation.

B. Limitations

Even though the suggested framework shows significant performance gains, there are still a number of drawbacks. First, function-level vulnerabilities are the main focus of the evaluation. Extended graph representations and inter-procedural analysis may be necessary to address vulnerabilities at the file or system level that involve interactions between multiple modules. Second, deep logical equivalency may not be fully captured by the semantic similarity metric because it depends on graph embeddings. For more robust assurances, future research might incorporate formal verification methods or symbolic execution. Third, the framework still relies on the caliber of test suites and static analyzers even though it minimizes false fixes. Subtle behavioral regressions could go unnoticed due to inadequate test coverage. Fourth, benchmark datasets that are accessible to the general public are used for the evaluation. More intricate vulnerability patterns and environment-specific limitations might be found in real-world industrial systems. Lastly, the explainability module does not yet offer formal proofs of correctness; instead, it produces structured reasoning based on validation signals. There is still room to improve explanation fidelity using logic-based reasoning.

VI. Conclusion

In order to improve automated vulnerability repair systems' dependability, transparency, and semantic integrity, this study

presented an Explainable LLM-Based Security Patch Validation Framework. The suggested framework combines test-driven validation, graph-based semantic consistency analysis, and static vulnerability verification into a single pipeline, in contrast to traditional LLM-driven techniques that only use generative confidence. Experiments on several benchmark datasets show notable gains in explanation fidelity, false-fix reduction, semantic preservation, and vulnerability removal accuracy. The framework outperforms both conventional and contemporary LLM-based repair systems, achieving a 96.3% vulnerability removal accuracy and a 9.3% reduction in false fixes. The results emphasize how crucial it is to combine structured validation methods with neural generation in order to accomplish reliable automated software repair. Future research will examine deployment in actual secure development pipelines, cross-language generalization, and integration with formal verification techniques. All things considered, the suggested framework is a workable and scalable step toward trustworthy and comprehensible AI-assisted security patch validation.

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 study and research output. All authors reviewed and approved the final manuscript.

ACKNOWLEDGMENT

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

REFERENCES

[1] H. Mazumdar, M. Sathvik, C. Chakraborty, B. Unhelkar, and S. Mahmoudi, "Real-time mental health monitoring for metaverse consumers to ameliorate the negative impacts of escapism and post trauma stress disorder," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2129–2136, 2024.

[2] H. Mazumdar, C. Chakraborty, M. Sathvik, P. K. Panigrahi *et al.*, "Gptfx: a novel gpt-3 based framework for mental health detection and explanations," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[3] A. Fan, B. Gokkaya, M. Harman, M. Lyubarskiy, S. Sengupta, S. Yoo, and J. M. Zhang, "Large language models for software engineering: Survey and open problems," in *2023 IEEE/ACM International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE)*. IEEE, 2023, pp. 31–53.

[4] C. Gao, X. Hu, S. Gao, X. Xia, and Z. Jin, "The current challenges of software engineering in the era of large language models," *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 5, pp. 1–30, 2025.

[5] S. Narkedimilli, S. Makam, A. V. Sriram, S. Prashanth Mallellu, M. Sathvik, and R. V. Prasad, "Enhancing iot network security through adaptive curriculum learning and xai," in *2025 IEEE Future Networks World Forum (FNWF)*, 2025, pp. 1–6.

[6] S. Han, H. Kim, H. Lee, H. Moon, Y. Jeon, H. Bae, D. Yeo, G.-J. Ahn, and S. Lee, "Dependable code repair with llms: Ai-driven vulnerability detection and automated patching," in *2025 IEEE 30th Pacific Rim International Symposium on Dependable Computing (PRDC)*. IEEE, 2025, pp. 176–181.

[7] Z. Sheng, Z. Chen, S. Gu, H. Huang, G. Gu, and J. Huang, "Llms in software security: A survey of vulnerability detection techniques and insights," *ACM Computing Surveys*, vol. 58, no. 5, pp. 1–35, 2025.

[8] S. Panichella, "Vulnerabilities introduced by llms through code suggestions," in *Large language models in cybersecurity: threats, exposure and mitigation*. Springer, 2024, pp. 87–97.

[9] M. Fu, J. Pasuksmit, and C. Tantithamthavorn, "Ai for devsecops: A landscape and future opportunities," *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 4, pp. 1–61, 2025.

[10] H. Yu, Y. Lai, L. Zhang, X. Lian, F. Liu, Y. Dong, T. Zhang, Z. Jin, and D. Lo, "Aligning academia with industry: An empirical study of industrial needs and academic capabilities in ai-driven software engineering," *arXiv preprint arXiv:2512.15148*, 2025.

[11] S. Latif, D. Djenouri, Z. Idrees, J. Ahmad, Z. Zou *et al.*, "Hardware security modules for secure communications in the industrial internet of things," *IEEE Communications Surveys & Tutorials*, 2025.

[12] K. Ahi and S. Valizadeh, "Large language models (llms) and generative ai in cybersecurity and privacy: A survey of dual-use risks, ai-generated malware, explainability, and defensive strategies," in *2025 Silicon Valley Cybersecurity Conference (SVCC)*. IEEE, 2025, pp. 1–8.

[13] N. O. Jaffal, M. Alkhanafseh, and D. Mohaisen, "Large language models in cybersecurity: A survey of applications, vulnerabilities, and defense techniques," *AI*, vol. 6, no. 9, p. 216, 2025.

[14] R. Özbay, M. Çelebi, and U. Yavanoğlu, "The ai-cybersecurity nexus: How large language models are reshaping threat intelligence and digital defense," *IEEE Access*, 2026.

[15] N. Amanmadov, J. Iskanderov, and T. Abdullayev, "Trustgraph: A heterogeneous gnn for dynamic zero-trust policy enforcement in microservices." *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 12, 2025.

[16] H. Yalamancheli, N. Shaik, N. Amanmadov, T. Abdullayev, and B. A. Brahmandam, "Bugblaze: An explainable framework for detecting latent defect ignitions in software systems," in *2025 IEEE 16th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2025, pp. 0644–0650.

[17] A. Bagheri, "Application of advanced ai methods for precise vulnerability detection," Ph.D. dissertation, University of Szeged, 2025.

[18] M. Alqarni, "Design and development of a vulnerability detection framework using artificial intelligence for embedded systems," 2025.

[19] N. T. Islam, G. D. L. T. Parra, D. Manual, M. Jadliwala, and P. Najafirad, "Causative insights into open source software security using large language code embeddings and semantic vulnerability graph," *arXiv preprint arXiv:2401.07035*, 2024.

[20] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, and H. Wang, "Large language models for cyber security: A systematic literature review," *ACM Transactions on Software Engineering and Methodology*, 2024.

[21] X. Zhou, S. Cao, X. Sun, and D. Lo, "Large language model for vulnerability detection and repair: Literature review and the road ahead," *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 5, pp. 1–31, 2025.

[22] I. Hasanov, S. Virtanen, A. Hakkala, and J. Isoaho, "Application of large language models in cybersecurity: A systematic literature review," *IEEE access*, vol. 12, pp. 176 751–176 778, 2024.

[23] J. Wang, T. Ni, W.-B. Lee, and Q. Zhao, "A contemporary survey of large language model assisted program analysis," *arXiv preprint arXiv:2502.18474*, 2025.

[24] Y. Wu, N. Jiang, H. V. Pham, T. Lutellier, J. Davis, L. Tan, P. Babkin, and S. Shah, "How effective are neural networks for fixing security vulnerabilities," in *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2023, pp. 1282–1294.

[25] R. Lin, Y. Fu, W. Yi, J. Yang, J. Cao, Z. Dong, F. Xie, and H. Li, "Vulnerabilities and security patches detection in oss: A survey," *ACM Computing Surveys*, vol. 57, no. 1, pp. 1–37, 2024.

[26] Q. Mao, Z. Li, X. Hu, K. Liu, X. Xia, and J. Sun, "Towards explainable vulnerability detection with large language models," *IEEE Transactions on Software Engineering*, 2025.

[27] N. Basheer, S. Islam, M. K. Alwaheidi, H. Mouratidis, and S. Papastergiou, "Large language model based hybrid framework for automatic vulnerability detection with explainable ai for cybersecurity enhancement," *Integrated Computer-Aided Engineering*, p. 10692509251368663, 2025.

[28] L. Belcastro, C. Carlucci, C. Cosentino, P. Liò, and F. Marozzo, "Enhancing network security using knowledge graphs and large language models for explainable threat detection," *Future Generation Computer Systems*, p. 108160, 2025.

[29] Y. Harrath, O. Adohinzin, J. Kaabi, and M. Saathoff, "Bridging domains: Advances in explainable, automated, and privacy-preserving ai for computer science and cybersecurity," *Computers*, vol. 14, no. 9, p. 374, 2025.

[30] A. A. Amoah and Y. Liu, "Explainable recommendation of software vulnerability repair based on metadata retrieval and multifaceted llms," *Machine Learning and Knowledge Extraction*, vol. 7, no. 4, p. 149, 2025.

[31] X. Huang, W. Ruan, W. Huang, G. Jin, Y. Dong, C. Wu, S. Bensalem, R. Mu, Y. Qi, X. Zhao *et al.*, "A survey of safety and trustworthiness of large language models through the lens of verification and validation," *Artificial Intelligence Review*, vol. 57, no. 7, p. 175, 2024.

[32] H. Mustafa, A. U. Soykat, R. Rahman, and M. B. Biplob, "A comprehensive review of large language models and ai in cybersecurity: Applications in threat detection, defense, and software security," 2025.

[33] H. Su, Z. Xu, Y. Zhang, and Q. Tan, "Source code vulnerability detection based on deep learning: a review," *Cybersecurity*, vol. 9, no. 1, p. 2, 2026.

[34] V. K. Anand, A. Das, and D. Chandawat, *Securing Large Language Models: Adversarial Attacks, Data Privacy, and Artificial Intelligence Safety*. Deep Science Publishing, 2026.

[35] A. Muhammed and R. Shekhar, "Unified artificial intelligence and formal reasoning approach to software security," in *2025 5th International Conference on Mobile Networks and Wireless Communications (ICMNWC)*. IEEE, 2025, pp. 1–7.

[36] N. I. of Standards and T. (NIST), "Juliet test suite for c/c++," 2017, available: https://samate.nist.gov/SARD/test-suites.

[37] Y. Zhou, S. Liu, J. Siow, X. Du, and Y. Liu, "Devign: Effective vulnerability identification by learning comprehensive program semantics," in *Advances in Neural Information Processing Systems*, 2019.

[38] R. Just, D. Jalali, and M. D. Ernst, "Defects4j: A database of existing faults to enable controlled testing studies for java programs," in *International Symposium on Software Testing and Analysis*, 2014.

[39] R.-M. Karampatsis and C. Sutton, "Manysstubs4j: A large dataset of java bug fixes for small and simple bugs," in *Mining Software Repositories*, 2020.

An Identity-Aware Privacy-Preserving Deep Learning Framework for Culturally Sensitive Image Sharing

Mahmoud Obaid*, Hadeel Bkhaitan, Duha Maali, Saja Hammad, Thaer Thaher
Department of Computer Systems Engineering, Arab American University, Jenin, P.O. Box 240, Palestine

***Abstract*—The blistering development of digital image sharing raises privacy concerns, especially in cultural contexts where image exposure could be ethically and socially provocative. In Islamic societies, sharing images of women without hijab can be deeply sensitive. This study presents SITR, an identity-sensitive privacy-preserving deep learning system aimed at reducing unintended sharing of sensitive images involving female family members without hijab. SITR integrates three components in a unified deployment-ready pipeline: 1) face recognition with Multi-task Cascaded Convolutional Networks (MTCNN), 2) family-member authentication with FaceNet embeddings stored in a vector database, and 3) hijab detection with an optimized Densely Connected Convolutional Network (DenseNet). The hijab detection model was trained and evaluated on a cleaned dataset of 2,191 images with hijab and non-hijab cases with diverse visual conditions. DenseNet121 was benchmarked against ResNet50, MobileNetV2, and EfficientNet-B0, achieving the best overall performance. To further enhance its effectiveness, DenseNet121 was modified by integrating an Efficient Channel Attention (ECA) mechanism and applying hyperparameter tuning. The optimized selected model achieved 92.16% test accuracy, strong discrimination with precision of 91.63% , and 86.39% F1-score on a held-out test set. The model was deployed as a quantized RESTful API, reduced from 82 MB to 27 MB while maintaining predictive reliability. Results demonstrate the practicability of identity-conditioned, culturally-aware AI systems for privacy protection. This work highlights the role of context-sensitive computer vision beyond generic content moderation toward culturally-aware and ethically accountable applications.**

***Keywords*—Cultural sensitivity; deep learning; DenseNet121; Efficient Channel Attention; FaceNet; hijab detection; MTCNN**

I. Introduction

In this age of digital interactions, transfer of photos is quick and immediate. Individuals post pictures to social networks, transfer them through communication services and archive them on cloud-based systems [1]. However, this feature raises the potential of unintended privacy exposure, especially in settings where personal images carry ethical, social, or cultural sensitivity. When a picture is posted on the internet, the creator of the image might never know who might see, save, or repost it [2]. Recent studies on image privacy protection further confirm that digital images often contain sensitive personal information and that protecting such content remains an active research challenge [3]. This concern is special in the Islamic world, where people appreciate modesty and personal privacy [4], [5].

The hijab is a head covering worn by women, and it is a sign of modesty and devotion to religion. For many women, images taken without a hijab are considered private and may lead to discomfort, social harm, or ethical concerns if shared unintentionally [6]. This risk is not limited to social media platforms. It also extends to messaging applications, email, and cloud storage services, where private images may be accidentally opened or even accidentally sent [7]. Such images are likely to be affected by human error when such protection is done manually, it is likely to go unnoticed when sharing such images and can be time-consuming as it requires checking and modifying.

In recent years, Artificial Intelligence (AI) techniques have been widely used in image analysis, face recognition, and content moderation [8], [9]. Photo management systems, security applications, and social media monitoring tools incorporate AI-powered technologies that can detect faces, organize photo collections, and remove inappropriate content. However, most of these systems treat such tasks independently and are not designed for privacy enforcement in culturally sensitive scenarios. Consequently, current tools are generally insufficient for situations in which sharing decisions depend not only on what appears in the image, but also on who appears in it, particularly to guard hijab-wearing women against unintentional exposure in the online space. Therefore, privacy-aware face-related systems remain an open research area.

Despite the fact that the independent research on face detection, face recognition or hijab detection has been conducted in several studies, there has been an imperative gap in identity-aware and culturally sensitive AI systems, which combine the contextual identity verification with the visual attribute classification into a unified, deployable framework. Current solutions are either general content moderation systems, hijab-related image classification approaches, or face-recognition systems without privacy-enforcement logic. This limits their use in culturally sensitive contexts where the perception of privacy cannot be solely based on the visual information in the image, but also on the identity of the people in the picture. In addition, computer vision under partial facial covering or appearance changes remains challenging, as shown in recent surveys on masked and occluded face analysis [10].

To address this gap, this study proposes SITR, an AI-powered system application that integrates deep learning models to help prevent the unintended sharing of privacy-sensitive photos. The system combines three deep learning modules: 1) face detection using MTCNN, 2) family member verification using FaceNet, and 3) hijab detection using a fine-tuned DenseNet121 model. When a female face is detected and identified as a registered family member, the system checks for the presence of a hijab. If no hijab is found, the face and neck are automatically blurred before sharing the photo. This design reflects the actual system workflow, where face

*Corresponding author

detection, identity verification, and hijab classification operate as a single privacy-aware pipeline before image sharing.

This study presents the design, development, and evaluation of the SITR system. The main contributions of this work are:

- A novel identity-conditioned privacy-preserving framework that integrates face detection, family-member verification, and hijab classification into a unified mobile pipeline.
- A systematically optimized hijab-detection component using DenseNet121 architecture enhanced with Efficient Channel Attention (ECA) and validated through controlled ablation.
- A deployment-aware implementation including model compression and RESTful API integration for real-world mobile usage.
- An empirical evaluation demonstrating the trade-off between predictive performance and computational efficiency in culturally sensitive AI applications.

The remainder of this study is organized as follows: Section II reviews related work. The system architecture and methodology are described in Section III. Section IV discusses the experimental results and system limitations. Finally, Section V concludes the study and outlines future work.

II. Related Work and Technical Background

This section explores the technical background, systems, and research studies relevant to SITR. It also presents the main deep learning models and techniques that support the SITR pipeline. No existing systems were identified that fully match the goals and architecture of SITR. However, several applications and studies partially address specific components relevant to the proposed framework, offering valuable insights that helped shape the system design.

A. CNN-Based Deep Learning and DenseNet

A Convolutional Neural Network (CNN) is a deep learning model that is widely used for working with visual data especially images. It uses filters to scan the input image and learn patterns such as edges, textures, and shapes. Because of this ability, CNNs have been successfully used in image classification, object detection, and face recognition. Based on CNNs, many deep learning models have been developed to improve accuracy and overall performance. These models keep the main idea of CNNs for learning features, but they also add new architectural improvements. DenseNet, MobileNet, EfficientNet, and ResNet are among the most common CNN-based models that are widely used in image classification and differ in their depth and complexity [11].

Dense Convolutional Network (DenseNet) is a deep learning architecture for CNNs introduced by Huang et al. [12]. It revolutionized the field of computer vision by proposing a novel connectivity pattern within CNNs, addressing challenges such as feature reuse, vanishing gradients, and parameter efficiency. Unlike traditional CNNs, each layer passes its output to all the following layers. This helps the network reuse features more effectively and improves the flow of information. As a result, important information is less likely to be lost as it moves through the network. DenseNet-121 is a version of the DenseNet architecture that has 121 layers, including convolutional, pooling, and fully connected layers. Because of this design, DenseNet-121 can achieve strong performance while using fewer parameters. Its efficient structure makes it a good choice for tasks that need deep feature extraction with lower computational cost [13].

B. Face Detection Using MTCNN

Face detection serves as a basic function of computer vision because it enables the detection and localization of human faces in digital images. Face-based applications usually start with this step, which supports later tasks such as facial recognition, emotion analysis, and identity verification. The Multi-task Cascaded Convolutional Network (MTCNN) developed by Zhang et al. functions as a deep learning system which detects multiple faces while finding their facial features. The system uses three cascading stages, including the Proposal Network (P-Net) and Refinement Network (R-Net) and Output Network (O-Net) [14]. MTCNN shows strong results on the WIDER FACE benchmark dataset because its cascaded design enables it to attain high accuracy and recall rates. MTCNN is used as the core method for detecting faces in user-uploaded images before further processing is applied.

C. Face Verification Using FaceNet Embeddings

Schroff et al. introduced FaceNet as a deep CNN model for face identification and verification, with the added capability of face clustering [15]. Rather than depending only on conventional classification, FaceNet projects face images into a compact Euclidean space where each face is represented by a discriminative embedding. These embeddings are learned using a triplet loss function, which encourages images of the same person to be placed closer together (anchor and positive samples) while forcing images of different people to remain farther apart (anchor and negative samples). Through training on large-scale datasets, FaceNet learns rich and robust facial features that allow it to perform effectively under variations in lighting, facial expression, and head pose. Another key strength of FaceNet is its ability to generate compact face embeddings, which makes storage and similarity comparison more efficient. This property makes the model well-suited for large-scale face recognition applications, particularly in mobile and cloud-based environments [16].

Within SITR, FaceNet is used to verify whether a detected face belongs to a known family member. By comparing the embedding of a new face to those stored in the database, the system can determine identity and proceed to hijab classification only for matched females.

D. Existing Related Studies and Systems

In this section, existing systems and related studies relevant to SITR are reviewed. No systems were identified that fully correspond to the vision and functionality of the proposed application; however, several platforms and research works address some of the targeted aspects and offer valuable insights that helped shape the proposed framework.

Ku and Dong [17] proposed a face recognition method based on the integration of MTCNN and a modified VGGNet architecture. Their model replaces max pooling layers with mean pooling and combines SoftMax Loss and Center Loss to enhance classification performance. The system achieves 98.53% accuracy on the Labeled Faces in the Wild (LFW) dataset. While this work shows strong performance in face recognition, it focuses on facial identification and does not include hijab detection or privacy-aware content protection.

In [5], a fully convolutional network (FCN) is used for automatic pixel-wise segmentation of faces, hijabs, and backgrounds. The model, trained on a set of 250 images, achieved 92.69% mean accuracy. This method simplifies hijab segmentation compared to manual approaches. However, the study focuses on segmentation and does not address identity verification, automated blurring, or mobile deployment. In another study, Khaliluzzaman et al. [18] introduced a visual feature-based hijab detection framework. Their method uses the Viola-Jones face detector and the YCbCr color model to identify hair and neck visibility. The approach achieved 96.47% accuracy on hijab images and 95% on non-hijab cases. It is important to note that this rule-based method operates under controlled visual conditions with explicitly visible neck and hair features, which may explain its higher reported accuracy compared to the deep learning baseline in this work. Deep learning approaches, however, offer greater generalizability and adaptability to unseen visual conditions, as demonstrated by the optimized model achieving 92.16% accuracy on a more heterogeneous dataset.

Other tools, such as HaramBlur, a browser extension, allows users to navigate the web with reduced browsing distractions by detecting and blurring "haram" content, including faces without hijab, using face detection and NSFWfiltering. However, it does not perform hijab classification or user verification. PimEyes [19], on the other hand, is a facial recognition search engine that allows users to locate images of a person online by uploading one photo. However, it lacks hijab detection and privacy-enforcement functionality. Table I presents a comparison of different face detection, authentication, and hijab detection systems alongside our proposed system, SITR, based on several key aspects.

E. Research Gap and Positioning of SITR

The reviewed systems and studies show meaningful progress in face detection, face recognition, and hijab-related image analysis. However, as shown in Table I, none of the reviewed systems provides a complete solution that combines face detection, face verification, and hijab detection, along with automated content blurring and mobile-friendly deployment. Most existing tools either focus on facial identification without considering cultural sensitivity or address hijab detection without verifying user identity.

This highlights the need for a privacy-focused tool designed for the specific needs of hijab-wearing users. SITR is positioned as an integrated mobile application that brings these components together in one pipeline. It is also worth noting that the proposed framework is specifically designed for the Islamic cultural context; its applicability to other cultural or religious contexts involving head coverings, such as medical caps, traditional headscarves in non-Islamic traditions, or sports equipment, is not evaluated in this work and represents a recognized scope limitation that should be addressed in future research.

III. System Overview and Methodology

This section describes the system design, architecture, and methodology of the proposed SITR application. The system integrates mobile and cloud components with deep learning models to enable privacy-aware image sharing. It also presents the operational workflow, hijab detection model, dataset preparation, training setup, and deployment process. Fig. 1 depicts an overview of the main processing pipeline used in the system. Overview of the SITR pipeline showing four sequential stages: 1) input image upload to the mobile application, 2) face detection localization using MTCNN, 3) identity verification via FaceNet embeddings and cosine similarity comparison against the ChromaDB database, and 4) hijab detection using the optimized DenseNet121 model. If a registered female family member is detected without a hijab, automatic face and neck blurring is applied before sharing. If a hijab is detected or the face does not match a registered member, it is safe to share.

A. System Architecture

SITR follows a modular architecture consisting of a mobile front-end, backend APIs, databases, and deep learning models. The architecture and data flow are illustrated in Fig. 2. As depicted in Fig. 2, the frontend layer communicates with the backend via HTTP REST calls. The backend orchestrates three sequential AI operations, including face detection (MTCNN), identity verification (FaceNet + ChromaDB), and hijab classification (DenseNet121), before returning a privacy enforcement decision to the mobile client.

The main system components are:

1) Frontend layer (mobile app): React Native and Expo are the used frameworks, and this layer will allow the interface to upload photos, manage family members, and see the results of detection. Support on user authentication and session management are provided to achieve secure access.

2) Backend/API layer: This layer is implemented using the Flask framework, which processes the image, along with the AI flow, by consuming RESTful APIs for face detection, face verification, and hijab classification.

3) Database layer: Convex stores structured user and family member information, while ChromaDB stores high-dimensional facial embeddings and enables real-time similarity queries using cosine similarity.

4) AI Engine layer: Includes three core models integrated in a sequential pipeline:

a) MTCNN: Detects faces and localizes facial landmarks with high accuracy under varied lighting and poses.

b) FaceNet: Generates 512-dimensional embeddings to verify detected faces against registered family members stored in ChromaDB.

c) DenseNet121-based hijab detection model: A fine-tuned DenseNet121 model used for hijab detection. ECA was later investigated as an enhancement during optimization.

TABLE I. COMPARISON OF RELATED SYSTEMS WITH SITR

System	Face Detection	Face Verification	Hijab Detection	Accuracy	Blurring Function	Remarks
MTCNN + Modified VGGNet [17]	Yes	Yes	No	98.53% (LFW)	No	Focused on face recognition using CNN with Center Loss.
FCN Segmentation Model [5]	No	No	Yes	92.69% (mean accuracy)	No	Pixel-wise segmentation into skin, hijab, and background.
Visual Feature Detection [18]	Yes	No	Yes	Hijab: 96.47%, Non-hijab: 95%	No	Rule-based detection using Viola-Jones and YCbCr color model.
HaramBlur Extension [20]	Yes	No	No	Not reported	Yes	Browser extension for NSFW content detection and blurring.
PimEyes Search Tool [19]	Yes	Yes	No	Not reported	No	Online facial search engine without cultural sensitivity.
SITR (Proposed)	Yes	Yes	Yes	92.16% (test set)	Yes	Integrates detection, verification, hijab classification, and blurring in a mobile application.

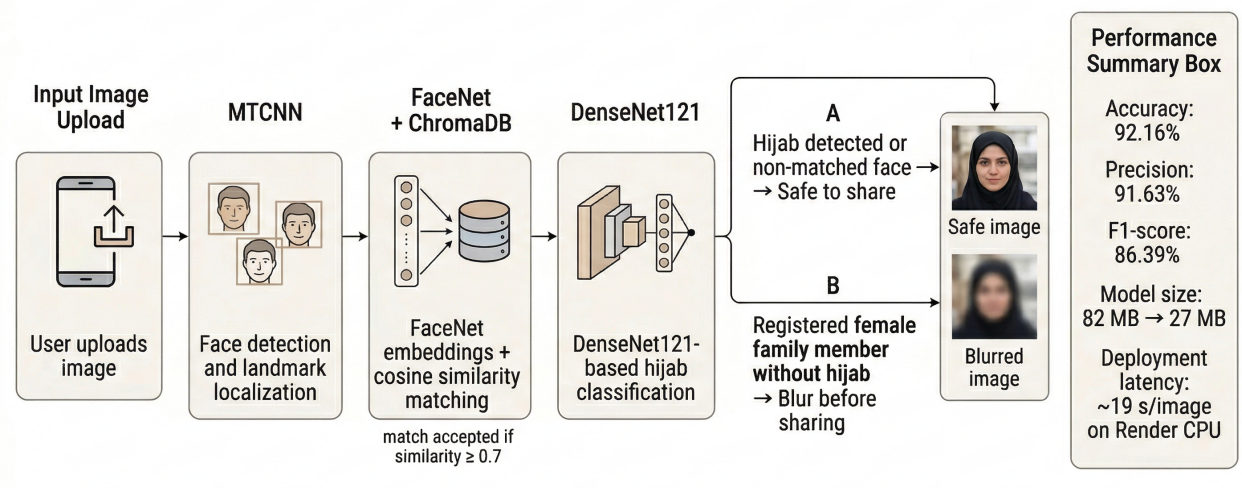


Fig. 1. Overview of the SITR pipeline: identity verification, hijab classification, privacy-enforcement decisions, and final deployment outcomes.

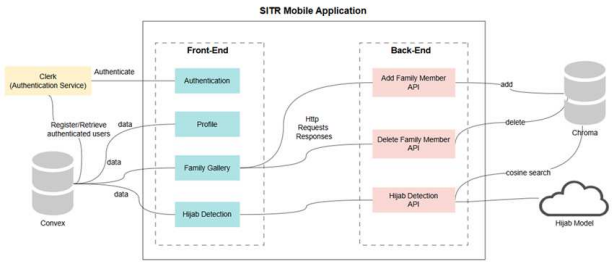


Fig. 2. System architecture with component interaction and data flow.

B. Operational Workflow

The operational workflow of SITR consists of two main stages, as illustrated in Fig. 3.

1) Enrollment stage: The user loads family member images into the application; the images are resized and passed through MTCNN for face detection, then through FaceNet to generate 512-dimensional facial embeddings, which are stored in ChromaDB for future matching.

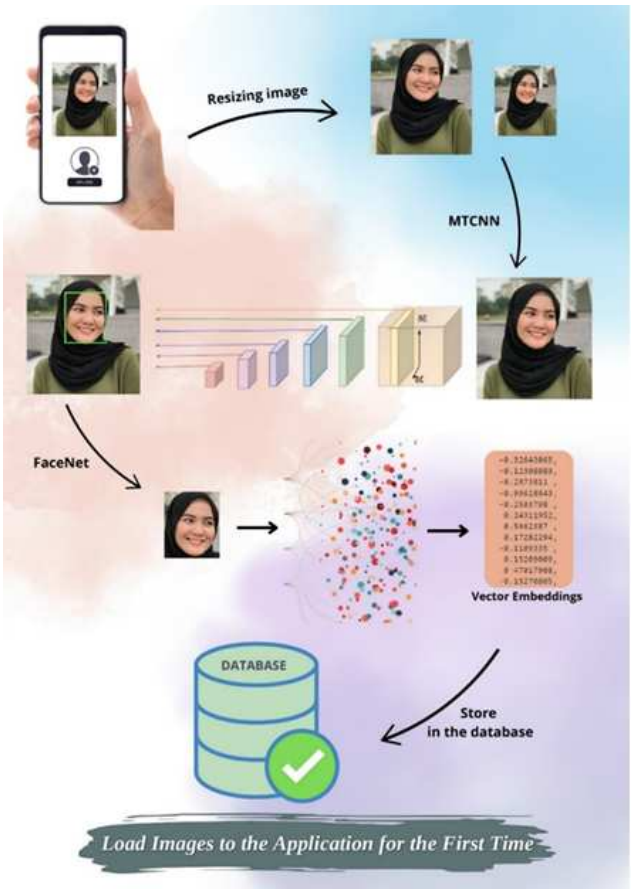
2) Detection stage: An uploaded image is resized, and MTCNN detects all faces. For each detected face, FaceNet generates an embedding that is compared against stored embeddings using cosine similarity (threshold $>= 0.7$). If a female family member is matched, the DenseNet121 hijab detection model classifies the region. If no hijab is detected, the face and neck region are automatically blurred before the image can be shared. Users are notified of the outcome and may choose to share or cancel.

C. Cosine Similarity for Face Matching

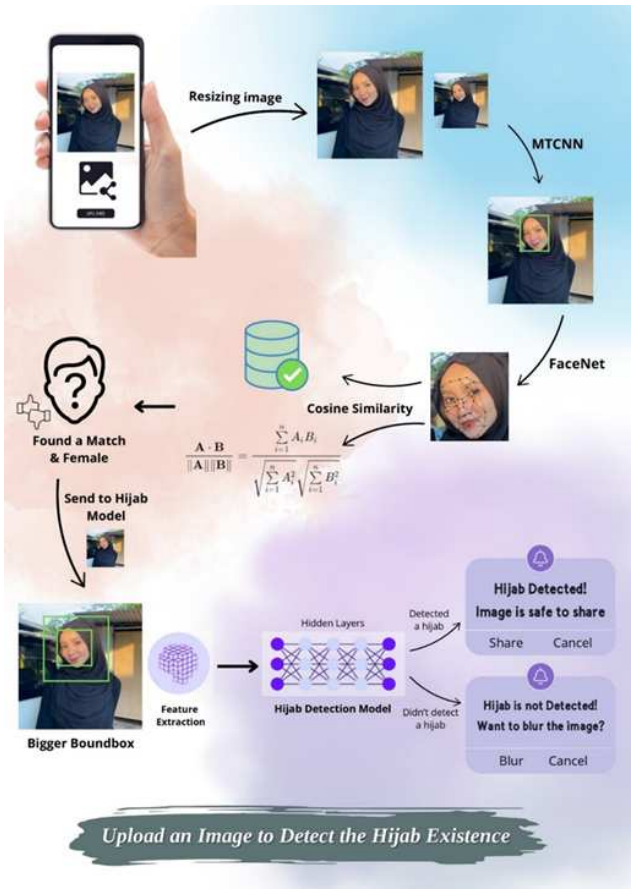
Once facial embeddings are generated using FaceNet, the next step in face verification involves comparing these embeddings to determine similarity. One of the most commonly used techniques for this comparison is cosine similarity. In SITR, this metric is used to verify whether a detected face belongs to a registered family member by comparing the embedding of the detected image with the stored face embeddings in the database. The cosine similarity between two embeddings is computed as follows:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

where, A and B are two face embeddings, $A \cdot B$ is the dot product, and $\|A\|$ and $\|B\|$ are the magnitudes of the vectors. A cosine similarity threshold of 0.7 was selected in SITR following a series of empirical trials in which thresholds



(a) Initial enrollment with embedding storage.



(b) hijab detection and privacy enforcement.

Fig. 3. SITR operational workflow including: (a) Family member enrollment and (b) Privacy-aware hijab detection.

ranging from 0.45 to 0.9 were evaluated on enrollment and test image pairs. The threshold of 0.7 yielded the best balance between true acceptance rate and false acceptance rate on the available face verification samples. It is acknowledged, however, that a formal ROC curve analysis and threshold sweep using a dedicated held-out verification set were not conducted; this represents a limitation of the current study and is recommended for future validation. Scores greater than or equal to 0.7 are accepted as a match, while scores below 0.7 are rejected.

D. Hijab Detection Model Based on DenseNet121

The hijab detection module of the SITR system was developed using DenseNet121 [12], a deep CNN pre-trained on the ImageNet dataset. This model was selected after evaluating multiple CNN architectures due to its robust feature extraction capabilities and proven effectiveness in image classification tasks. DenseNet121 is well known for its strong performance because it captures fine-grained features through dense blocks, where each layer is connected to every other layer in a feed-forward fashion [21]. This design promotes feature reuse and effective combination [22], which makes it suitable for detecting small or subtle details, such as whether a female subject is wearing a hijab. Since it is pre-trained on a large and diverse dataset, DenseNet121 provides a strong foundation for this classification task and adapts well to varied lighting conditions and viewing angles, allowing high accuracy even with a relatively small dataset [13].

To further improve its effectiveness, additional optimization steps were investigated:

1) Efficient Channel Attention (ECA): ECA [23], a light and effective channel-attention module based on local cross-channel interaction via 1D convolution, was added after each dense block to help the model concentrate on the most informative channel features.

2) Hyperparameter tuning: A grid search [24] was conducted to determine optimal values for learning rate, dropout, batch size, and weight decay.

The customized DenseNet121 architecture used for hijab detection is illustrated in Fig. 4. It shows the complete modified DenseNet121 architecture. The input image (224×224×3 px) passes through an initial 7×7 convolution and 3×3 max pooling layer before entering four dense blocks. After each dense block, an ECA attention layer is inserted for lightweight channel-wise feature recalibration using a 1D convolution (kernel size k=3). Transition layers (1×1 convolution + 2×2 average pooling) reduce spatial dimensions between dense blocks. After the fourth Dense Block, a 7×7 global average pooling layer is applied, followed by a dropout layer (rate=0.3)

and a fully connected linear layer that produces the binary classification output (0 = no hijab; 1 = hijab). The effect of these optimization steps is evaluated later in the Results and Discussion section.

E. Dataset and Preprocessing

A dataset of 3,082 annotated images collected from Roboflow was used to train the hijab detection model. Following data cleaning, 891 samples were removed due to inconsistent annotation or poor image quality. The final valid dataset contained 2,191 images (including 1,248 training images, 63 validation images, and 880 testing images). There was no bias due to class imbalance as there was a balance between the hijab and non-hijab classes (50.8% vs. 49.2, respectively). All the splits were done using a fixed seed randomly to guarantee reproducibility. The test set was not used in the hyperparameter tuning and was strictly held out. It is acknowledged that this dataset size is relatively small for training a robust deep learning classifier. In particular, the validation split of only 63 images may limit the reliability of hyperparameter selection decisions. These constraints stem from the limited availability of publicly annotated, culturally appropriate data and represent a recognized limitation of this study. Future work should aim to collect larger and more diverse datasets incorporating varied geographic regions, hijab styles, and demographic backgrounds to improve model generalizability. Cross-dataset validation is also recommended to assess robustness beyond the current data distribution.

F. Training and Evaluation Setup

The hijab detection model was trained and tested on Google Colab with GPU support, which ensured faster and more efficient experiments. All CNN models were trained under consistent conditions for fairness, as presented in Table II. The Colab environment was equipped with 2 vCPUs, 12 GB RAM, and an NVIDIA T4 GPU with 16 GB VRAM. Moreover, default PyTorch parameters were used unless otherwise noted. To improve reliability, four independent runs were conducted for each model, and results were averaged.

TABLE II. TRAINING SETUP AND PARAMETERS USED FOR ALL CNN MODELS

Factor	Value
Image Input Size	224×224 px
Batch Size	4
Learning Rate	1e-5
Weight Decay	1e-6
Dropout	0.3
Epochs	20
Optimizer / Loss	Adam with binary cross-entropy loss

G. Evaluation Metrics

To evaluate the effectiveness and reliability of the SITR hijab detection model, several commonly used metrics in classification tasks were adopted, as summarized in Table III. All metrics were calculated from multiple independent runs to reduce the impact of random variations in model performance and provide a more trustworthy assessment of the model's effectiveness. All the reported results are the averages over four independent runs.

TABLE III. SUMMARY OF EVALUATION METRICS USED TO ASSESS SITR MODEL.

Metric	Description / Purpose
Accuracy	Percentage of correctly classified images (hijab / non-hijab).
Precision	Fraction of predicted "hijab" images that are actually correct.
Recall	Fraction of actual hijab images correctly detected by the model.
F1-score	Harmonic mean of precision and recall for balanced evaluation.
Computational Time	Training and inference time, reflecting model efficiency.
Loss Values	Indicates training stability and convergence during evaluation.

H. Deployment Setup

To support practical use of the optimized DenseNet121 model, it was deployed as a REST API using the Flask framework on the Render cloud platform. Dynamic quantization reduced the model size to 27 MB from the original 82 MB, showing the memory efficiency and speed benefits. The SITR mobile application uses the backend API to communicate with Convex for handling user data and with ChromaDB to support similarity search during family-member verification.

IV. RESULTS AND DISCUSSION

This section presents quantitative and qualitative results for evaluating the SITR application and its core hijab detection model. The experiment was performed through three different phases: initially by benchmarking CNN models, followed by optimization of the DenseNet121 model with Efficient Channel Attention (ECA), and finally by tuning hyperparameters for maximum performance.

A. Quantitative Results

1) Benchmarking advanced CNN models: To identify the baseline model for hijab detection, four pre-trained CNN architectures, including ResNet50 [25], MobileNetV2 [26], EfficientNet-B0 [27], and DenseNet121) were investigated on the curated dataset. These models are selected because they are widely used for image classification and represent different levels of depth and complexity [13].

Table IV summarizes the average performance metrics across four separate runs. Fig. 5 presents grouped bar charts comparing accuracy, precision, recall, and F1-score across the four architectures, with each bar representing the mean value across four independent training runs to allow assessment of both peak performance and training consistency. The results show that DenseNet121 achieved the best overall performance. It recorded an average accuracy of 79.49%, precision of 73.48%, recall of 62.69%, and an F1-score of 65.67%. These findings represent a clear improvement over other model. Its accuracy is 5.3% higher than MobileNetV2, 3.9% higher than ResNet50, and 0.8% higher than EfficientNet-B0. In terms of F1-score, DenseNet121 also outperformed its closest competitor, EfficientNet-B0, by 1.1%. It should be noted that the baseline DenseNet121 accuracy of 79.49% is lower than the 96.47% reported for the rule-based visual feature detection method proposed in [18]. This discrepancy is

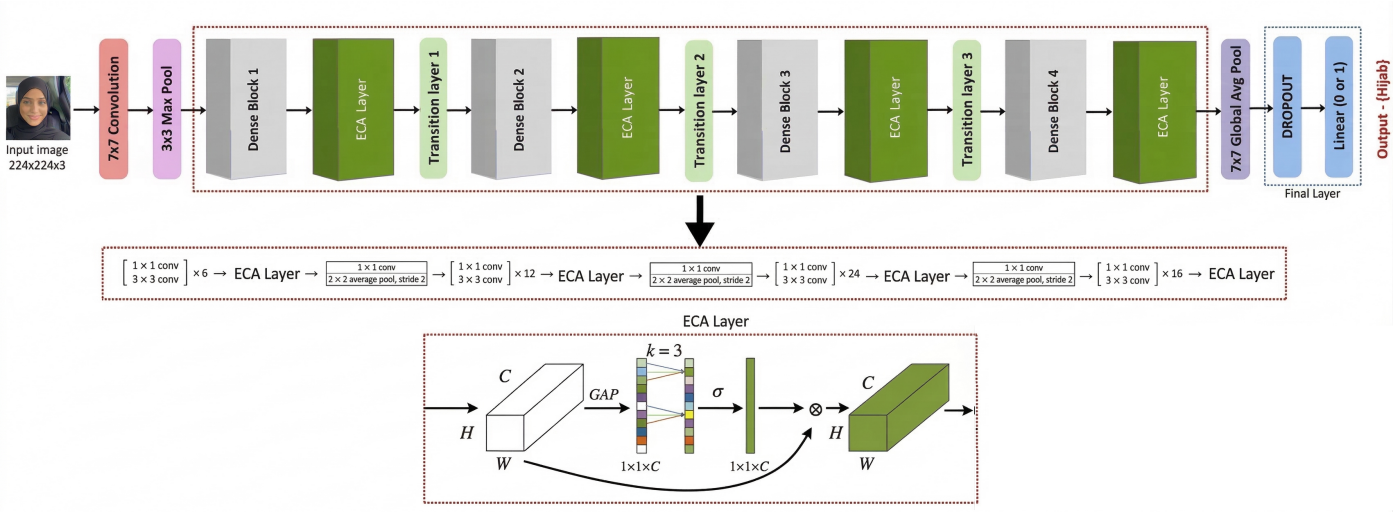


Fig. 4. The customized DenseNet121 architecture used for hijab detection model.

primarily attributable to differences in evaluation conditions: the rule-based approach was tested under controlled settings with clearly visible neck and hair features, whereas the deep learning models in this study were evaluated on a more heterogeneous dataset encompassing diverse lighting conditions, poses, partial occlusions, and backgrounds. Deep learning approaches such as DenseNet121 are expected to generalize better to unseen real-world conditions, as evidenced by the improved accuracy of 92.16% achieved after optimization, while rule-based methods may degrade significantly under uncontrolled visual conditions.

TABLE IV. AVERAGE PERFORMANCE METRICS ACROSS FOUR RUNS ON THE TEST SET IN TERMS OF CLASSIFICATION METRICS AND LOSS.

Model	Classification Metrics (%)				Loss
	Accuracy	Precision	Recall	F1-Score	
ResNet50	75.63	60.91	77.89	66.52	0.4512
MobileNetV2	73.18	55.87	60.36	57.84	0.4541
EfficientNet-B0	78.69	67.63	60.73	64.95	0.481
DenseNet121	79.49	73.48	62.69	65.67	0.486

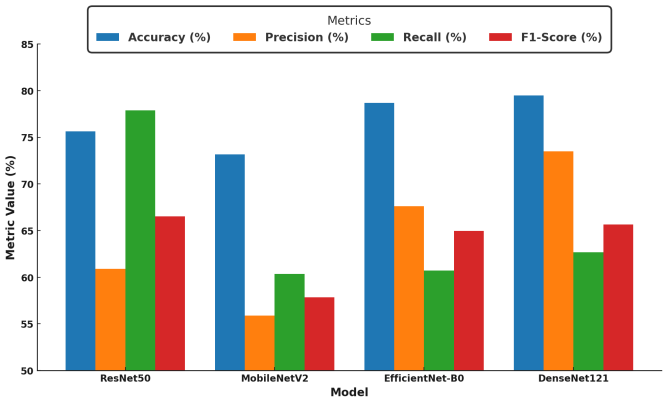


Fig. 5. Test classification results for the four models (averaged over four runs).

Precision is a critical metric in this context because it reflects the model's ability to reduce false positives. DenseNet121 achieved the highest precision at 73.48%, which indicates greater reliability in distinguishing hijab from non-hijab images. While ResNet50 captured slightly more true hijab samples, its lower precision reduced its overall consistency. In contrast, DenseNet121's architecture uses dense connectivity for better feature reuse. This allowed it to capture fine-grained details of hijab texture and coverage, leading to more stable and accurate classification.

Fig. 6 depicts the training and validation loss for each model. It is clear that DenseNet121 shows the smoothest and most consistent convergence. Its training and validation losses decreased steadily and remained close to each other. This pattern indicates strong generalization and reliable optimization. EfficientNet-B0 also achieves low final losses. However, its validation curve fluctuated during the middle epochs, which suggests it adapted more slowly to variations in the data. Meanwhile, validation loss for MobileNetV2 declines less uniformly This observation is consistent with its lower accuracy and recall scores in Table IV. ResNet50 reaches a reasonable loss value early in training, but its validation loss stabilizes at a higher level compared to DenseNet121. Overall, DenseNet121 converged faster and reached the lowest combined losses, supporting its superior quantitative results.

2) Model optimization with ECA: After selecting DenseNet121 as the best-performing baseline, it was further enhanced by integrating Efficient Channel Attention (ECA) after each dense block to prioritize informative channels. As shown in Table V, the ECA-enhanced version achieves an accuracy of 86.62% and an F1-score of 77.32%. This marks a significant improvement over the standard DenseNet121, which records 79.49% accuracy and a 65.67% F1-score. Overall, accuracy increases by 7.13% and the F1-score by 11.65%.

Other metrics improve as well. Precision rises from 73.48% to 79.91%, and recall increases from 62.69% to 74.91%. These findings confirm that the attention mechanism creates a better balance between identifying true hijab samples

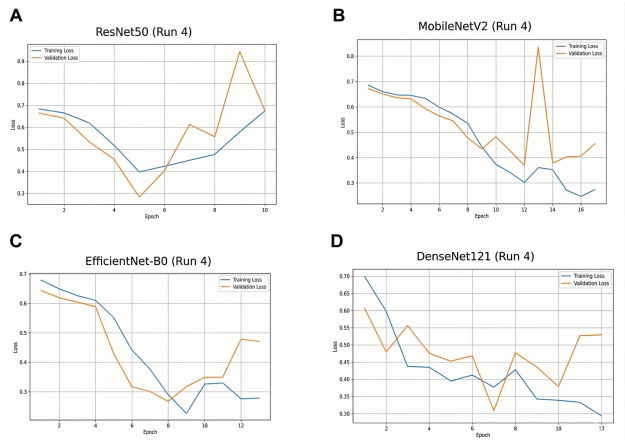


Fig. 6. Training and validation loss curves of four tested models over multiple epochs (Run 4).

and reducing false detections. The lower loss value, 0.369 compared to 0.486, also indicates more stable convergence. Together, these results demonstrate that incorporating ECA helps DenseNet121 learn richer feature representations and improves its overall robustness in hijab detection. However, this comparison reflects improvement over the untuned standard DenseNet121 only.

TABLE V. COMPARISON OF STANDARD AND ECA-ENHANCED DENSENET121 MODELS (AVERAGE OVER 4 RUNS).

Variant	Classification Metrics (%)				Loss
	Accuracy	Precision	Recall	F1-Score	
Standard DenseNet121	79.49	73.48	62.69	65.67	0.486
ECA-Enhanced	86.62	79.91	74.91	77.32	0.369

3) Hyperparameter tuning: A grid search was performed across various learning rates, dropout rates, batch sizes, and weight decays. The hyperparameter and the optimal configuration are given in Table VI. Due to GPU limitations, 19 different combinations were evaluated.

TABLE VI. GRID SEARCH HYPERPARAMETERS AND OPTIMAL VALUES

Parameter	Search Space	Optimal
Learning Rate	[0.001, 0.0001, 0.00001]	0.00001
Dropout	[0.3, 0.45, 0.7]	0.3
Batch Size	[4, 8, 16]	4
Weight Decay	[0.0001, 0.00001, 0.000001]	0.000001

Although the ECA-enhanced variant performed exceptionally well, it did not surpass the regular DenseNet121 after it was properly tuned. As such, we chose the tuned DenseNet121 as our final model on the basis of its superior validation performance in terms of accuracy, F1-score, and overall generalization capability.

The final deployment model was selected based on validation performance after hyperparameter tuning and then evaluated on the held-out test set. It achieved a test accuracy of 92.16%, an F1-score of 86.39%, and a test loss of 0.2174. It also achieved a precision of 91.63%. This high-precision value is particularly significant for the privacy-enforcement context of SITR. A false positive — where a non-hijab image is misclassified as containing a hijab — results in the system incorrectly declaring a privacy-sensitive image safe to share, representing the most critical failure mode. The achieved precision of 91.63%, therefore, reflects a low false-positive rate, directly aligned with the system's primary objective of protecting user privacy before image sharing.

B. Qualitative Results

To verify practical system behavior, the SITR mobile application was tested for image upload, family member verification, hijab detection, and privacy enforcement. During the enrollment stage, family members were successfully registered using FaceNet embeddings, which supported reliable verification in later stages. In the photo-sharing stage, uploaded images were processed and the system automatically verified hijab status and applied blurring to sensitive faces when necessary. These examples illustrate the intended behavior of the SITR pipeline in real image-sharing scenarios.

As illustrated in Fig. 7, the system correctly identified hijabi family members and allowed their images to be safely shared. The left screenshot shows the application in the analysis phase, where MTCNN detects the face and FaceNet confirms the identity as a registered female family member with a cosine similarity above the 0.7 threshold. The center screenshot displays the result: the DenseNet121 model detects a hijab and clears the image for sharing, displaying a "You can share the photo safely" confirmation message. The right screenshot shows the upload interface after analysis, indicating that no blurring was applied. This scenario corresponds to Path A in the SITR pipeline. In contrast, Fig. 8 shows how the system automatically blurred the head and neck region when a registered female family member was detected without a hijab, thereby enforcing privacy protection before the image could be shared. These examples highlight SITR's real-world functionality in supporting privacy-sensitive image sharing.

C. Discussion on Model Efficiency and Limitations

The strong performance of DenseNet121 can be attributed to its dense connectivity mechanism, which enables effective feature reuse and preserves low-level texture information across layers—characteristics particularly well-suited to hijab detection, which requires recognition of subtle visual patterns including fabric coverage, hair visibility, and neck exposure. The introduction of ECA improves the channel-wise feature recalibration and enables the network to focus on semantically valuable channels, and also reduces redundant responses. The mechanism is especially useful in visually complicated backgrounds where noise in the context may disrupt classification.

The optimized model was successfully deployed using Render after applying dynamic quantization. This technique reduces the model size from 82 MB to 27 MB. This compression ratio demonstrates that the proposed system is well-suited for resource-constrained deployment environments, including mobile and edge devices. The findings suggest that identity-aware, culturally sensitive AI systems of this nature are practically feasible beyond the laboratory setting and can be integrated as lightweight API services into existing mobile applications. An

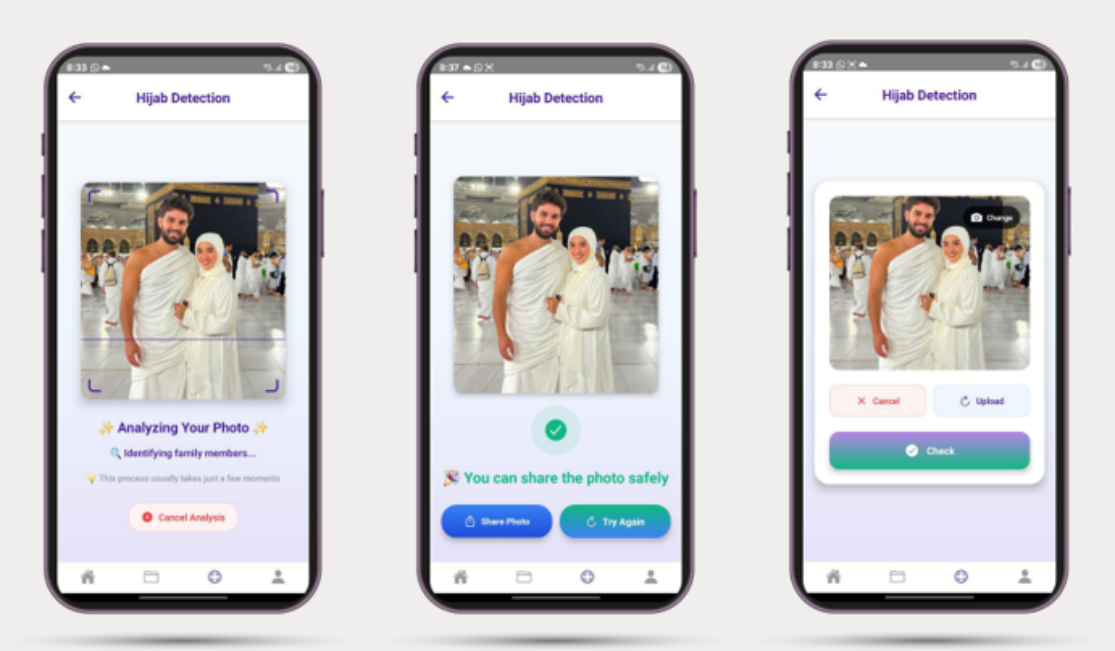


Fig. 7. Example of a hijabi family member correctly identified as safe for sharing.

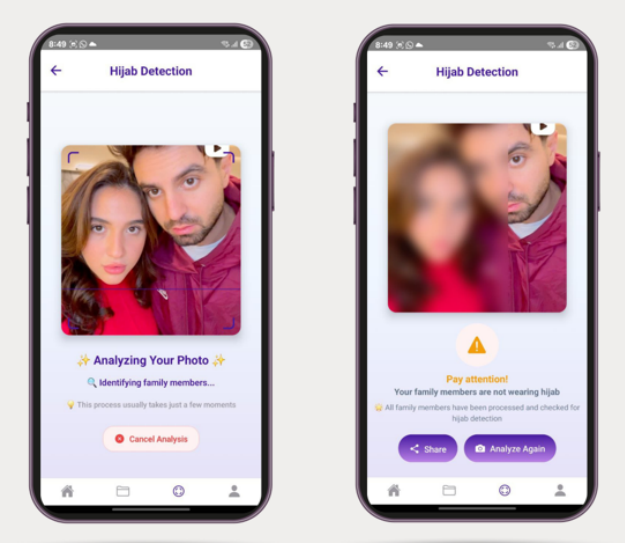


Fig. 8. Example of automatic blurring applied to a family member not wearing a hijab.

average inference time of approximately 19 seconds per image was recorded on a CPU-only cloud instance. This latency is acceptable for privacy-sensitive manual image-sharing workflows but is not yet suitable for real-time streaming applications.

From a broader research perspective, SITR demonstrates the viability of combining identity verification with visual attribute classification in a culturally grounded AI pipeline. This design pattern may be extended to other culturally sensitive contexts - such as age-appropriate content filtering or faith-based social media moderation—although such extensions would require separate dataset collection, model training, and ethical evaluation.

Regarding the absence of an end-to-end system evaluation, the current work evaluates each pipeline component (MTCNN, FaceNet, DenseNet121) individually. No integrated pipeline-level evaluation was conducted to measure the compounded effect of errors across all three stages on the final privacy decision output. A missed face detection or an incorrect identity match may propagate errors that are not captured by the individual component metrics reported in this study. This is a recognized limitation, and future work should include end-to-end evaluation using realistic multi-face test scenarios to quantify the true system-level accuracy.

Despite strong results, the system has several notable limitations:

1) Small and limited dataset: The hijab detection model was trained on only 2,191 images after removing 891 samples for quality issues. The validation split of only 63 images further limits the reliability of hyperparameter selection. This small and geographically constrained dataset may hinder generalization to diverse populations, hijab styles, and demographic backgrounds.

2) No end-to-end pipeline evaluation: Each of the three pipeline components — face detection, identity verification, and hijab classification — was evaluated independently. The compounded error across the full pipeline in realistic multi-face image sharing scenarios remains unmeasured.

3) Restricted cultural scope: The framework is explicitly designed for the Islamic cultural context. It does not account for women who wear head coverings for non-religious reasons (e.g., medical head coverings, traditional or cultural dress outside the Islamic context), which could produce misclassifications. The system is not validated for other cultural or religious head-covering traditions.

V. Conclusion

This study proposed the SITR framework, a culturally-aware and identity-sensitive deep learning system designed to prevent unintended sharing of privacy-sensitive images, specifically targeting the Islamic cultural context of hijab wearing. In contrast to traditional content moderation systems, SITR bases decisions on privacy enforcement on confirmed identity as well as classification in visual attributes in a single pipeline. The combination of face detection, identity verification through facial embeddings, and optimized hijab recognition shows how AI can be implemented in context-sensitive ethical use. The final tuned DenseNet121 model achieved an accuracy of 92.16% and an F1-score of 86.39%. Experimental results showed that ECA improved the untuned baseline significantly, but the tuned standard DenseNet121 achieved the best final performance and was selected for deployment. Dynamic quantization reduced the model size from 82 MB to 27 MB, making it lightweight and suitable for REST API deployment. The system demonstrated smooth end-to-end communication between the mobile frontend and backend, supporting practical integration in privacy-sensitive image-sharing scenarios.

Despite these promising results, two primary limitations must be explicitly acknowledged. First, the hijab detection model was trained on a relatively small dataset (2,191 images) with a minimal validation split (63 images), which limits confidence in the generalizability of the learned representations to diverse cultural, demographic, and environmental contexts. Second, no end-to-end system evaluation was conducted to quantify the compounded error across the three-stage pipeline (face detection $\rightarrow$ identity verification $\rightarrow$ hijab classification), meaning the true system-level accuracy in realistic multi-face image-sharing scenarios remains unmeasured. Additionally, the cosine similarity threshold of 0.7 used for face verification was selected empirically without a formal ROC-based analysis, which represents a further methodological gap.

Future work should address these gaps through the following directions: 1) collecting larger, more culturally and demographically diverse hijab detection datasets and applying cross-dataset validation; 2) conducting a full end-to-end pipeline evaluation with realistic test scenarios to measure true system-level precision and error propagation; 3) performing formal ROC curve analysis for threshold selection in the face verification component; and 4) exploring GPU-optimized and lighter model architectures to reduce inference latency below the current approximately 19 seconds per image, enabling real-time or near-real-time privacy enforcement in dynamic image-sharing contexts.

Declaration on Generative AI

During the preparation of this manuscript, the authors used a generative AI tool (ChatGPT) only to assist with language refinement, paraphrasing, and text organization. After using this tool, all generated content was carefully reviewed and edited by the authors as needed. The authors remain fully responsible for the scientific content, accuracy, originality, and integrity of this manuscript.

Acknowledgment

The authors would like to thank the Department of Computer Systems Engineering at the Arab American University for supporting the senior project that formed the basis of this work. The authors also acknowledge all contributors who supported the implementation, testing, and refinement of the SITR system.

References

[1] K. Ghazinour and J. Ponchak, "Hidden privacy risks in sharing pictures on social media," *Procedia Computer Science*, vol. 113, pp. 267–272, 2017.

[2] B. Debatin, J. P. Lovejoy, A.-K. Horn, and B. N. Hughes, "Facebook and online privacy: Attitudes, behaviors, and unintended consequences," *Journal of Computer-Mediated Communication*, vol. 15, pp. 83–108, 2009.

[3] L. Rakhmawati, Wirawan, and Suwadi, "Image privacy protection techniques: A survey," in *TENCON 2018 - 2018 IEEE Region 10 Conference*, 2018, pp. 0076–0080.

[4] T. Afnan, Y. Zou, M. Mustafa, M. Naseem, and F. Schaub, "Aunties, strangers, and the FBI: Online privacy concerns and experiences of Muslim-American women," Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022). USENIX Association, 2022, pp. 387–406.

[5] A. Madani and M. W. F. D. M. Madkour, "Automatic face and hijab segmentation using convolutional network," *International Journal of Integrated Engineering*, vol. 11, pp. 60–66, 2019.

[6] T. Afnan, Y. Zou, M. Mustafa, M. Naseem, and F. Schaub, "Aunties, strangers, and the fbi: online privacy concerns and experiences of muslim-american women," in *Proceedings of the Eighteenth USENIX Conference on Usable Privacy and Security*, ser. SOUPS'22. USA: USENIX Association, 2022.

[7] M. Franco, S. A. Falioun, K. E. Fisher, O. Gaggi, Y. Ghamri-Doudane, A. J. Nashwan, C. E. Palazzi, and M. Shwamra, "A technology exploration towards trustable and safe use of social media for vulnerable women based on islam and arab culture," in *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, ser. GoodIT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 138–145. [Online]. Available: https://doi.org/10.1145/3524458.3547259

[8] T. Gillespie, "Content moderation, ai, and the question of scale," *Big Data & Society*, vol. 7, 2020.

[9] W. Ali, W. Tian, S. U. Din, D. Iradukunda, and A. A. Khan, "Classical and modern face recognition approaches: a complete review," *Multimedia Tools Appl.*, vol. 80, no. 3, p. 4825–4880, Jan. 2021. [Online]. Available: https://doi.org/10.1007/s11042-020-09850-1

[10] K. M. Hosny, N. AbdElFattah Ibrahim, E. R. Mohamed, and H. M. Hamza, "Artificial intelligence-based masked face detection: A survey," *Intelligent Systems with Applications*, vol. 22, p. 200391, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667305324000668

[11] S. Ram, S. Vinoth, R. N. Gopalakrishnan, A. A. Balakumar, L. Kalinathan, and T. A. J. Velankanni, "Leveraging diverse CNN architectures for medical image captioning: DenseNet-121, MobileNetV2, and ResNet-50 in ImageCLEF 2024," in *CLEF 2024 Working Notes, Proceedings of the 15th International Conference of the CLEF Association*, ser. CEUR Workshop Proceedings, vol. 3740. Grenoble, France: CEUR-WS.org, 2024, pp. paper–160. [Online]. Available: https://ceur-ws.org/Vol-3740/paper-160.pdf

[12] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[13] T. Thaher, M. Saffarini, M. Mafarja, A. Alashbi, A. H. Mohamed, and A. A. El-Saleh, "A hybrid approach for heavily occluded face detection using histogram of oriented gradients and deep learning models," *Computer Modeling in Engineering & Sciences*, vol. 144, no. 2, pp. 2359–2394, 2025. [Online]. Available: http://www.techscience.com/CMES/v144n2/63698

[14] Z. L. Z. Zhang and Y. Q. K. Zhang, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, pp. 1499–1503, 2016.

[15] D. Kalenichenko and J. P. F. Schroff, "Facenet: A unified embedding for face recognition and clustering." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.

[16] A. Zhalgas, B. Amirgaliyev, and A. Sovet, "Robust face recognition under challenging conditions: A comprehensive review of deep learning methods and challenges," *Applied Sciences*, vol. 15, no. 17, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/15/17/9390

[17] W. K. H., & Dong, ""face recognition based on mtcnn and convolutional neural network."," *Frontiers in Signal Processing, vol. 4, no. 1*, pp. 37–42, 2020.

[18] A. Begum, S. K. Fatama, and M. M. I. M. Khaliluzzaman, "Detection and analysis of hijab based on visual feature of neck and hair," *in 2017 20th International Conference of Computer and Information Technology (ICCIT)*, pp. 1–6, Dec. 2017.

[19] Pimeyes - advanced face recognition search engine. [Online]. Available: Available: https://pimeyes.com/en

[20] A. Aljanzory. (2024) Haramblur - blur haram, nsfw images & videos. [Online]. Available: Available: https://chromewebstore.google.com/detail/haramblur-blur-haram-nsfw/pbcoegikffnadpahojjhgdladmmddeji?hl=en

[21] P. Podder, F. B. Alam, M. R. H. Mondal, M. J. Hasan, A. Rohan, and S. Bharati, "Rethinking densely connected convolutional networks for diagnosing infectious diseases," *Computers*, vol. 12, no. 5, 2023. [Online]. Available: https://www.mdpi.com/2073-431X/12/5/95

[22] F. A. Azis, H. Suhaimi, and E. Abas, "Real-time face mask classification with convolutional neural network for proper and improper face mask wearing," *International Journal of Computing*, vol. 22, no. 2, pp. 184–190, Jul. 2023. [Online]. Available: https://computingonline.net/computing/article/view/3087

[23] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 531–11 539.

[24] J. Ilemobayo, O. Durodola, O. Alade, O. Awotunde, T. Adewumi, O. Falana, A. Ogungbire, A. Osinuga, D. Ogunbiyi, I. Odezuligbo, O. Edu, and A. Ifeanyi, "Hyperparameter tuning in machine learning: A comprehensive review," *Journal of Engineering Research and Reports*, vol. 26, pp. 388–395, 06 2024.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[27] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds. PMLR, 2019, pp. 6105–6114. [Online]. Available: http://proceedings.mlr.press/v97/tan19a.html

Stability-Weighted Feature Selection with Adaptive PSO-SGD for Neural Network-Based Predictive Hiring

Yassine Temsamani Khallouk, Said Achchab
Mohammed V University in Rabat, Morocco

***Abstract*—Predictive hiring leverages machine learning to forecast candidate success, yet existing approaches suffer from two limitations: reliance on single-method feature selection that lacks robustness, and sensitivity to neural network initialization that impairs convergence. This study introduces two contributions integrated into a unified framework. First, Stability-Weighted Multi-Criteria Feature Selection (SW-MCFS) is proposed, which aggregates four heterogeneous scoring methods—Mutual Information, Wald statistical significance, Fisher discriminant loading, and Permutation Importance—through a cross-validation stability-weighted consensus function. Unlike single-method approaches, SW-MCFS weights each method proportionally to its ranking consistency across folds, producing robust and data-driven feature subsets. Second, Adaptive Particle Swarm Optimization (APSO) is introduced, a PSO variant featuring fitness-landscape-aware inertia adaptation and Lévy flight perturbation for stagnation escape. The framework is evaluated on $10,247$ recruitment records from a North African telecommunications company and benchmarked against Random Forest, XGBoost, SVM, standard ANNs, and classical LR/DA-based approaches through 10-fold cross-validation. The integrated SW-MCFS-APSO-SGD framework achieves 76.8% accuracy, significantly outperforming XGBoost (73.8%, $p = 0.012$), standard PSO-SGD (75.2%, $p = 0.041$), and LR-based feature selection (74.6%, $p = 0.028$). Ablation studies confirm that SW-MCFS contributes 1.6% accuracy gain over single-method selection, while APSO improves performance by 0.8% with 31% faster convergence compared to standard PSO. SHAP analysis reveals communication skills, experience, and seniority as dominant predictors with minimal demographic influence. It is noted that the accuracy ceiling may partly reflect inherent label noise in subjective performance assessments. The proposed framework demonstrates effectiveness on organizational recruitment data, warranting further cross-domain validation to establish broader generalizability.**

***Keywords*—Predictive hiring; Multi-Criteria Feature Selection; Adaptive Particle Swarm Optimization; stability-weighted consensus; neural networks; HR analytics; ensemble methods**

I. Introduction

A. Background and Motivation

Modern recruitment processes generate large volumes of structured and unstructured data, including resumes, assessments, interviews, and work history. Traditional hiring practices struggle to analyze this information objectively and at scale, resulting in inefficiencies, bias, and suboptimal talent selection [1], [2], [30]. Predictive hiring applies machine learning to forecast candidate success by leveraging historical employment data to identify performance-related patterns [3], [32].

Recent advances in artificial intelligence (AI) have enabled automated candidate evaluation through natural language processing (NLP) for resume analysis [4], psychometric profiling [5], and automated interview assessment [6].

B. AI-Driven Transformation of Human Resource Information Systems

AI integration transforms Human Resource Information Systems (HRIS) from transactional platforms into intelligent decision-support systems. Traditional HRIS mainly manages payroll, personnel records, and compliance [7]. With AI integration, these systems evolve into advanced analytics-driven infrastructures [8], [9].

AI technologies, including ML, deep learning (DL), NLP, and large language models (LLM), enable HRIS to process heterogeneous and large-scale workforce data. This supports evidence-based human resource management and strengthens core functions including talent acquisition, workforce planning, and performance management [10]. Strategically, AI enables HR systems to transition toward intelligent and data-driven workforce management with enhanced analytical capabilities.

C. AI Techniques for Recruitment Automation and Predictive Modeling

NLP techniques such as tokenization, semantic embedding, and named entity recognition enable automated extraction of structured features from textual recruitment data [11], [12]. Transformer-based architectures [42] improve semantic similarity estimation and candidate–job matching through contextual representation learning.

LLM further enhances automation by improving feature representation and supporting intelligent screening and classification [13]. Predictive hiring systems then use these extracted features to estimate post-hire performance using ML models. This study integrates advanced feature selection and optimization methods to improve predictive robustness within AI-driven HR environments.

D. Research Challenges

Despite recent advances, two fundamental challenges remain:

- *Feature Selection Fragility:* Most existing studies rely on a single feature selection technique such as correlation filtering, LR significance testing, or tree-based

importance ranking. These approaches often produce unstable feature subsets that vary across data partitions and model assumptions [14]. However, no consensus-based framework systematically aggregates multiple selection criteria while accounting for stability across cross-validation folds.

- *Optimization Sensitivity:* Standard Particle Swarm Optimization (PSO) employs fixed or predefined inertia weight schedules that do not adapt to the fitness landscape topology [15]. When used to optimize neural network (NN) parameters, this limitation may result in premature convergence or insufficient exploration, reducing predictive performance and generalization ability.

E. Research Contributions

To address these challenges, this study makes the following contributions:

1) Stability-Weighted Multi-Criteria Feature Selection (SW-MCFS): A feature selection framework is proposed that aggregates multiple heterogeneous scoring methods, including (1) Mutual Information (MI), (2) Wald Statistical Significance (WSS), (3) Fisher Discriminant Loading (FDL), and (4) Permutation Importance (PI), into a unified consensus ranking. Each method is weighted based on its cross-validation stability to enhance robustness. To the best of the authors' knowledge, this is the first stability-weighted multi-criteria consensus framework applied to predictive hiring feature selection.

2) Adaptive PSO (APSO): An adaptive PSO variant is introduced incorporating (1) fitness-based inertia adaptation, (2) swarm diversity monitoring, and (3) Lévy flight perturbations to enhance exploration and prevent premature convergence during NN weight optimization.

3) Integrated hybrid learning framework: An end-to-end pipeline is developed that integrates SW-MCFS with adaptive APSO for initialization and stochastic gradient descent (SGD) for training. This joint design improves feature robustness and model optimization within a unified framework that has not been explored in prior predictive hiring studies.

4) Rigorous empirical validation: Comprehensive evaluation on $10,247$ records using 10-fold cross-validation, ablation analysis, statistical tests, and baseline comparisons to demonstrate robustness and effectiveness.

F. Paper Organization

The remainder of this study is organized as follows: Section II reviews related work and theoretical foundations. Section III describes the proposed optimization framework. Section IV presents the experimental setup. Section V and Section VI reports the results and discussion. Finally, Section VII concludes the study and outlines future research directions.

II. Related Work

A. Feature Selection in HR Analytics

Feature selection in HR prediction typically include filter methods (e.g., correlation, MI), wrapper methods (e.g., recursive feature elimination), and embedded methods (e.g., L1 regularization, tree-based importance) [16]. Most studies rely on a single technique such as Logistic Regression (LR) significance testing [17] or Random Forest (RF) importance [18].

However, these approaches often suffer from instability and bias. Filter methods ignore feature interactions, wrapper methods are computationally expensive, and embedded methods depend on the base model [19]. Although ensemble feature selection has been explored in other domains [20], stability-aware consensus strategies remain limited in HR prediction. The proposed SW-MCFS framework addresses this gap by incorporating stability-weighted aggregation.

B. PSO Variants for NN Optimization

PSO is widely used for NN optimization due to its global search ability [15]. Standard PSO uses fixed inertia strategies, which limit adaptability to complex landscapes.

Adaptive variants improve exploration through dynamic inertia control [21], state-based adaptation [22], or Lévy flight perturbations [23]. However, existing methods rarely combine fitness-aware adaptation with Lévy-based exploration for predictive hiring models. The proposed APSO variant integrates these mechanisms to improve optimization stability.

C. AI in Predictive Hiring

ML for recruitment has evolved from traditional classifiers to deep learning models. NN-based person–job fit models achieve competitive accuracy [24], [33], while co-attention architectures and ensemble methods, including gradient boosting variants such as XGBoost [26] and CatBoost [41] further improve performance [25].

Despite these advances, feature selection and model training are usually treated separately. Integrated optimization within a unified pipeline remains underexplored.

D. Research Gaps

Current literature lacks: 1) stability-weighted multi-criteria feature selection for HR prediction, 2) adaptive swarm optimization combining fitness-aware control and Lévy perturbation for neural training, and 3) unified frameworks that jointly optimize feature robustness and model learning. These gaps motivate the proposed SW-MCFS-APSO-SGD framework.

III. Proposed SW-MCFS–APSO Hybrid Optimization Framework

This section presents the proposed SW-MCFS–APSO–SGD framework, depicted in Fig. 1, that integrates stability-aware feature selection, adaptive optimization, and gradient-based training for robust predictive modeling.

A. Problem Formulation and Cross-Validation Framework

Let the training dataset be defined as Eq. (1):

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}, \tag{1}$$

where,

- N denotes the number of recruitment records,

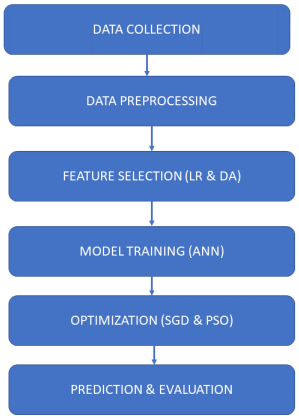


Fig. 1. SW-MCFS-APSO-SGD framework. Phase 1: Multi-criteria feature selection with stability weighting. Phase 2: APSO global search with FLAI, LFP, and EGR. Phase 3: SGD fine-tuning with early stopping.

- $\mathbf{x}_i = \{x_{i,j}\}_{1\le j\le d} \in \mathbb{R}^d$ is the raw feature vector of candidate i $(i \le N)$,
- d is the number of original features,
- $y_i \in \{0, 1\}$ is the binary job performance label, with $i \le N$.

Accordingly, the feature matrix can be expressed as Eq. (2):

$$\mathbf{M} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} \in \mathbb{R}^{N\times d}. \tag{2}$$

Each row corresponds to a candidate, and each column corresponds to a feature. To evaluate feature stability and prevent data leakage, K-fold cross-validation (CV) is adopted on $\mathcal{D}$. For each $k \le K$, the dataset $\mathcal{D}$ is partitioned into training and validation set as Eq. (3):

$$\mathcal{D} = \mathcal{D}_{\text{tr}}^{(k)} \cup \mathcal{D}_{\text{val}}^{(k)}, \quad 1 \le k \le K, \tag{3}$$

such that for each fold k [see Eq. (4)]:

$$\begin{cases} \mathcal{D}_{\text{tr}}^{(k)} \cap \mathcal{D}_{\text{val}}^{(k)} = \emptyset \\ \left|\mathcal{D}_{\text{val}}^{(k)}\right| = \lfloor \frac{N}{K} \rfloor, \quad \left|\mathcal{D}_{\text{tr}}^{(k)}\right| = N_k \triangleq N - \lfloor \frac{N}{K} \rfloor \end{cases}. \tag{4}$$

B. Multi-Criteria Feature Selection

Initially, the proposed SW-MCFS algorithm is applied to reduce the original feature set to a selected subset $\mathcal{F}$, where $d^* = |\mathcal{F}|$ denotes the number of selected features satisfying $d^* \le d$.

The reduced feature matrix becomes [see Eq. (5)]:

$$\mathbf{M}_{\mathcal{F}} \in \mathbb{R}^{N\times d^*}. \tag{5}$$

To achieve a robust and stable ranking, multiple complementary feature relevance criteria are aggregated. In particular, **four** scoring methods are employed to evaluate feature importance from statistical, information-theoretic, and model-driven perspectives.

1) Mutual information: For each feature $x_{i,j}$ (the j-th feature of sample i), the MI quantifies its dependency with the binary label y_i by measuring the reduction in uncertainty of y_i given $x_{i,j}$.

Since features are continuous, the joint and marginal distributions are estimated using kernel density estimation or discretization. The MI score for feature j is computed in fold k, as in [27] [see Eq. (6)]:

$$\mathcal{M}_j^{(k)} = \sum_{x,y} p_{X_j,Y}^{(k)}(x,y) \log \frac{p_{X_j,Y}^{(k)}(x,y)}{p_{X_j}^{(k)}(x)\, p_Y^{(k)}(y)}, j \le d, k \le K \tag{6}$$

where, X_j and Y denote the j-th feature and label RVs, respectively, and the joint distribution $p_{X_j,Y}^{(k)}(x,y)$ and the marginals $p_{X_j}^{(k)}(x)$ and $p_Y^{(k)}(y)$ are estimated empirically from $\mathcal{D}_{\text{tr}}^{(k)}$. Of note, a larger $\mathcal{M}_j^{(k)}$ implies stronger relevance of feature j to the target label.

For continuous features, $\mathcal{M}_j^{(k)}$ is approximated using the k-nearest-neighbor estimator [28]:

$$\widehat{\mathcal{M}}_j^{(k)} = \psi(k_{\text{nn}}) + \psi(N_k) - \frac{1}{N_k} \sum_{i=1}^{N_k} \Big[\psi(n_x^{(i)}+1) + \psi(n_y^{(i)}+1)\Big], \tag{7}$$

where, $n_x^{(i)}$ and $n_y^{(i)}$ denote the number of samples (excluding the i-th one) whose distances to sample i in the X_j and Y marginal spaces are smaller than the k_{nn}-nearest neighbor (NN) radius ε_i, which is determined by the distance to the k_{nn}-th NN in the joint space.

Lastly, to make scores comparable across features, the normalized MI score is defined based on the estimator in Eq. (7) as [see Eq. (8)]:

$$S_{j,k}^{(1,n)} = \frac{\widehat{\mathcal{M}}_j^{(k)}}{\max_{1\le \ell\le d} \widehat{\mathcal{M}}_\ell^{(k)}}, \quad 1 \le k \le K. \tag{8}$$

2) Wald statistical significance: For each fold, a univariate L2-regularized LR model is fitted to evaluate the statistical significance of feature X_j to the binary label Y. The conditional probability is modeled as Eq. (9):

$$\Pr(Y = 1 \mid X_j = x_{i,j}) = \frac{1}{1 + \exp\big(-\beta_{0,j}^{(k)} - \beta_{1,j}^{(k)} x_{i,j}\big)}. \tag{9}$$

The parameters $\beta_{0,j}^{(k)}$ and $\beta_{1,j}^{(k)}$ are estimated from $\mathcal{D}_{\text{tr}}^{(k)} \triangleq \{(x_{i,j}, y_i)\}_{i=1}^{N_k}$ via penalized maximum likelihood with L2 regularization to ensure numerical stability.

The null hypothesis $\mathcal{H}_0 : \beta_{1,j}^{(k)} = 0$ is tested using the Wald statistic [see Eq. (10)]:

$$W_j^{(k)} = \frac{(\hat{\beta}_{1,j}^{(k)})^2}{\widehat{\mathrm{Var}}(\hat{\beta}_{1,j}^{(k)})} \sim \chi^2(1), \qquad 1 \le j \le d, \tag{10}$$

where, $\chi^2(1)$ denotes the chi-square distribution with one degree of freedom, and $\widehat{\mathrm{Var}}(\hat{\beta}_{1,j}^{(k)})$ is obtained from the inverse Fisher information matrix.

The corresponding p-value is Eq. (11):

$$p_j^{(k)} = \Pr\big(\chi^2(1) \ge W_j^{(k)}\big) = 1 - F_{\chi^2(1)}(W_j^{(k)}), \tag{11}$$

where, $F_{\chi^2(1)}(\cdot)$ denotes the cumulative distribution function.

Similarly to Eq. (8), the normalized Wald-based score is defined as Eq. (12):

$$S_{j,k}^{(2,n)} = \frac{|\hat{\beta}_{1,j}^{(k)}|(1-p_j^{(k)})}{\max_{1\le\ell\le d}|\hat{\beta}_{1,\ell}^{(k)}|(1-p_\ell^{(k)})}. \tag{12}$$

This formulation jointly rewards large effect sizes and high statistical confidence while penalizing unstable estimates.

3) Fisher discriminant score: From Linear Discriminant Analysis [44], the discriminant direction that maximizes class separation in fold k is given by:

$$\mathbf{w}^{(*,k)} = (\mathbf{S}_W^{(k)})^{-1}(\boldsymbol{\mu}_1^{(k)} - \boldsymbol{\mu}_2^{(k)}), \tag{13}$$

where, $\mathbf{S}_W^{(k)}$ denotes the within-class scatter matrix computed on $\mathcal{D}_{\text{tr}}^{(k)}$ and $\boldsymbol{\mu}_c^{(k)}$ represents the mean vector of class c in fold k.

Each component $w_j^{(*,k)}$ of the discriminant direction in Eq. (13) measures the contribution of feature j to class separation. To obtain a scale-invariant measure, it is weighted by the pooled within-class standard deviation $\sigma_j^{(k)}$, and the normalized Fisher score in fold k is defined as Eq. (14):

$$S_{j,k}^{(3,n)} = \frac{|w_j^{(*,k)}|\,\sigma_j^{(k)}}{\max_{1\le\ell\le d}|w_\ell^{(*,k)}|\,\sigma_\ell^{(k)}}, \qquad 1 \le j \le d, \tag{14}$$

This normalization ensures comparability across features and highlights those with stronger discriminative power.

4) Permutation importance: Permutation importance is computed within each cross-validation fold using a baseline RF model with $B = 200$ trees [18], [29]. It quantifies the decrease in predictive performance when feature values are randomly permuted.

Let $f_\theta^{(k)}(\cdot)$ denote the classifier trained on the training subset $\mathcal{D}_{tr}^{(k)}$ in fold k. The accuracy functional evaluated on a dataset $\mathcal{D}$ in fold k is defined as Eq. (15):

$$\mathcal{A}^{(k)}(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{(\mathbf{x}_k,y_k)\in\mathcal{D}} \mathbf{1}\big(f_\theta^{(k)}(\mathbf{x}_k) = y_k\big), \tag{15}$$

where, $\mathbf{1}(\cdot)$ is the indicator function.

The baseline accuracy in fold k is computed on the original validation set as Eq. (16):

$$\mathcal{A}_{\text{original}}^{(k)} = \mathcal{A}^{(k)}(\mathcal{D}_{\text{val}}^{(k)}). \tag{16}$$

To evaluate the importance of feature j in fold k, its values are randomly permuted across the validation samples, generating a modified dataset $\pi_r(\mathcal{D}_{\text{val}}^{(k)}, X_j)$. The permutation importance score is defined as the average reduction in accuracy [see Eq. (17)]:

$$\mathcal{P}_j^{(k)} = \mathcal{A}_{\text{original}}^{(k)} - \frac{1}{R}\sum_{r=1}^{R}\mathcal{A}^{(k)}\Big(\pi_r(\mathcal{D}_{\text{val}}^{(k)}, X_j)\Big); 1 \le j \le d, \tag{17}$$

where, R denotes the number of random permutations used in each fold.

Finally, the normalized importance score is given by Eq. (18):

$$S_{j,k}^{(4,n)} = \frac{\mathcal{P}_j^{(k)}}{\max_{1\le\ell\le d}\mathcal{P}_\ell^{(k)}}; 1 \le j \le d. \tag{18}$$

To further improve robustness, SW-MCFS assigns stability-driven weights to the four methods, as detailed in the subsequent subsection, thereby reducing the influence of methods exhibiting inconsistent rankings across folds.

C. Stability-Based Weighting

To quantify the reliability of each method, its ranking stability is measured as the average pairwise Spearman correlation between fold-specific rank vectors:

$$\rho_m = \frac{2}{K(K-1)}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K}\rho_S\big(\mathbf{r}_m^{(k)}, \mathbf{r}_m^{(l)}\big), \tag{19}$$

where, Eq. (20):

$$\mathbf{r}_m^{(k)} = \big(r_{m,1}^{(k)}, \ldots, r_{m,d}^{(k)}\big), \quad m \le 4, 1 \le k \le K, \tag{20}$$

denote the rank vector of the d original features, and the Spearman rank correlation is:

$$\rho_S\big(\mathbf{r}_m^{(k)}, \mathbf{r}_m^{(l)}\big) = 1 - \frac{6\sum_{j=1}^{d}\big(r_{m,j}^{(k)} - r_{m,j}^{(l)}\big)^2}{d(d^2-1)}. \tag{21}$$

Substituting Eq. (21) into Eq. (19) yields Eq. (22):

$$\rho_m = 1 - \frac{12}{K(K-1)d(d^2-1)}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K}\sum_{j=1}^{d}\big(r_{m,j}^{(k)} - r_{m,j}^{(l)}\big)^2. \tag{22}$$

Each method is assigned a stability weight using a softmax normalization:

$$\omega_m = \frac{\exp(\gamma \rho_m)}{\sum_{\ell=1}^{4} \exp(\gamma \rho_\ell)}, \quad 1 \le m \le 4, \tag{23}$$

where, $\gamma > 0$ controls the sensitivity to stability differences. Larger values emphasize the most stable method, whereas smaller values lead to nearly equal weights (i.e., $\exp(\gamma \rho_m)/4$).

Leveraging Eq. (23), the final stability-weighted consensus score is defined as Eq. (24):

$$\overline{\mathcal{S}}_{j,w} = \sum_{m=1}^{4} \omega_m \, \bar{S}_j^{(m,n)}, \tag{24}$$

where, $\bar{S}_j^{(m,n)}$ is the average over the K folds of the mth score, assessed using Eq. (8), Eq. (12), Eq. (14), and Eq. (18) as [see Eq. (25)]:

$$\overline{S}_j^{(m,n)} = \frac{1}{K} \sum_{k=1}^{K} S_{j,k}^{(m,n)}, \quad 1 \le m \le 4. \tag{25}$$

It is noteworthy that:

- Since $\omega_m \ge 0$, $\sum_{m=1}^{4} \omega_m = 1$, and $\bar{S}_j^{(m)} \in [0,1]$, it follows that $\overline{\mathcal{S}}_{j,w} \in [0,1]$;
- Methods exhibiting low ranking stability (small ρ_m) receive smaller weights ω_m through the softmax normalization, thereby reducing their influence on the final consensus score.

Subsequently, features are ranked according to $\overline{\mathcal{S}}_{j,w}$. Denote $\bar{S}_{\alpha_j,w}$ the jth greatest final stability-weighted score, i.e., Eq. (26):

$$\bar{S}_{\alpha_1,w} \ge \bar{S}_{\alpha_2,w} \ge \cdots \ge \bar{S}_{\alpha_d,w}, \quad \{\alpha_j\}_{j \le d} \in \{1,\ldots,d\}. \tag{26}$$

Lastly, the number of retained features d^* is determined using an elbow-based adaptive threshold [see Eq. (27)]:

$$d^* = \min\{j \in \{2,\ldots,d-1\} : |\Delta \bar{S}_{\alpha_j,w}| > \eta\}, \tag{27}$$

where,

$$\Delta \bar{S}_{\alpha_j,w} = \bar{S}_{\alpha_{j+1},w} - 2\,\bar{S}_{\alpha_j,w} + \bar{S}_{\alpha_{j-1},w}, \tag{28}$$

and $\eta > 0$ is a small tolerance parameter [see Eq. (28)].

Algorithm 1 summarizes the aforementioned steps of the proposed SW-MCFS process, which produces the top-d^* features from the selected subset $\mathcal{F}$. In practice, the parameter γ is fixed to 5 in all experiments.

Algorithm 1 SW-MCFS: Stability-Weighted Multi-Criteria Feature Selection

Input: $K, N, d, \mathcal{D}, \gamma, \eta, R$
Output: $\mathcal{F}^*$
$\mathcal{C} \leftarrow \lfloor \frac{N}{K} \rfloor$ //Cardinal of $\mathcal{D}_{\text{val}}^{(k)}$
for $k = 1$ to K **do**
//Split $\mathcal{D}$ into $\mathcal{D}_{\text{val}}^{(k)}$ and $\mathcal{D}_{\text{tr}}^{(k)}$ using (3)-(4)
  $\mathcal{D}_{\text{val}}^{(k)} \leftarrow \{(\mathbf{x}_{(k-1)\times\mathcal{C}+i}, y_i)\}_{i=1}^{\mathcal{C}}$
  $\mathcal{D}_{\text{tr}}^{(k)} \leftarrow \mathcal{D} \setminus \mathcal{D}_{\text{val}}^{(k)}$
  for $m = 1$ to 4 **do**
    for $j = 1$ to d **do**
      Evaluate $S_{j,k}^{(m,n)}$ // using Eqs. (8),(12),(14),(18)
    end for
    Compute $\overline{S}_j^{(m,n)}$ //using Eq. (25)
    Deduce $\mathbf{r}_m^{(k)}$ //defined in Eq. (20)
  end for
end for
//Stability Computation
for $m = 1$ to 4 **do**
  Evaluate ρ_m and ω_m //Using Eq. (19) and (23)
end for
for $j = 1$ to d **do**
  Evaluate $\bar{S}_{j,w}$ // Using Eq. (24)
end for
//Adaptive Threshold
Sort features in descending order of $\{\bar{S}_{j,w}\}_{1 \le j \le d}$
$j = 2$
repeat
  Compute $\Delta \bar{S}_{\alpha_j,w}$ //using Eq. (28)
  $j \leftarrow j + 1$
until $\Delta \bar{S}_{\alpha_{j-1},w} \le \eta$ **and** $j \le d$
$d^* \leftarrow j - 1$
$\mathcal{F}^* = \{\alpha_1, \ldots, \alpha_{d^*}\}$
return $\mathcal{F}^*$

D. Adaptive PSO (APSO)

After SW-MCFS reduces the original feature space to a lower-dimensional space of size $d^* \le d$, optimization is performed using a swarm of $\mathcal{S}$ particles. To balance exploration and exploitation adaptively, the concept of normalized swarm diversity is introduced within the APSO framework. The following notation formally defines the optimization procedure.

1) Notation: At iteration $t = 1, \ldots, \mathcal{T}_{\text{APSO}}$, particle $i = 1, \ldots, \mathcal{S}$ is characterized by:

- $\mathbf{p}_i^{(t)} \in \mathbb{R}^{d^*}$: *position vector* of particle i, representing a candidate set of model parameters.
  It is initialized as Eq. (29):
$$\mathbf{p}_i^{(0)} \sim \mathcal{U}(-p_{\max}, p_{\max})^{d^*}. \tag{29}$$
- $\mathbf{v}_i^{(t)} \in \mathbb{R}^{d^*}$: *velocity vector* controlling the update direction and exploration strength. It is initialized as Eq. (30):
$$\mathbf{v}_i^{(0)} \sim \mathcal{U}(-v_{\max}, v_{\max})^{d^*}. \tag{30}$$
- $\mathbf{p}_{i,\text{best}}^{(t)}$: personal best position of particle i, defined as the best position visited up to iteration t:
$$\mathbf{p}_{i,\text{best}}^{(t)} = \arg\min_{\tau \le t} f(\mathbf{p}_i^{(\tau)}), \qquad 1 \le i \le \mathcal{S}.$$
- $f(\mathbf{p})$: *fitness function* evaluating the validation performance of a particle [see Eq. (31)],
$$f(\mathbf{p}) = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(x,y) \in \mathcal{D}_{\text{val}}} \mathcal{L}(F(x;\mathbf{p}), y), \tag{31}$$

where, $F(x;\mathbf{p})$ denotes the NN parameterized by $\mathbf{p}$, and $\mathcal{L}$ is the loss function measuring the discrepancy between the predicted output $\hat{y} = F(x;\mathbf{p})$ and the true label y.

- $\mathbf{g}^{(t)}$: *global best position* among all particles [see Eq. (32)],

$$\mathbf{g}^{(t)} = \arg\min_{1\le i\le \mathcal{S}} f(\mathbf{p}_{i,\text{best}}^{(t)}). \tag{32}$$

The overall optimization problem is formulated as Eq. (33):

$$\mathbf{p}^* = \arg\min_{\mathbf{p}\in\mathbb{R}^{d^*}} f(\mathbf{p}). \tag{33}$$

2) Fitness-Landscape-Aware Inertia (FLAI): Instead of using a fixed inertia weight, it is adapted according to the current behavior of the swarm [21], [22].

The improvement of the global best at iteration t is defined as Eq. (34):

$$\Delta^{(t)} = \frac{f(\mathbf{g}^{(t-1)}) - f(\mathbf{g}^{(t)})}{f(\mathbf{g}^{(t-1)}) + \varepsilon}, \tag{34}$$

where, ε avoids numerical instability.

Swarm diversity is quantified by the dispersion of personal best positions [see Eq. (35)]:

$$\mathcal{F}^{(t)} = \frac{1}{\mathcal{S}\times d^*} \sum_{i=1}^{\mathcal{S}} \|\mathbf{p}_{i,\text{best}}^{(t)} - \bar{\mathbf{p}}_{\text{best}}^{(t)}\|_2, \tag{35}$$

where [see Eq. (36)],

$$\bar{\mathbf{p}}_{\text{best}}^{(t)} = \frac{1}{\mathcal{S}} \sum_{i=1}^{\mathcal{S}} \mathbf{p}_{i,\text{best}}^{(t)}. \tag{36}$$

Using the normalized versions $\hat{\mathcal{F}}^{(t)}$ and $\hat{\Delta}^{(t)}$ defined as Eq. (37):

$$\hat{\chi}^{(t)} = \frac{\chi^{(t)} - \chi_{\min}^{(t)}}{\chi_{\max}^{(t)} - \chi_{\min}^{(t)}}, \quad \chi \in \{\mathcal{F}, \Delta\}, \tag{37}$$

where Eq. (38),

$$\chi_{\min}^{(t)} = \min_{k\in[t-W+1,\,t]} \chi^{(k)}, \quad \chi_{\max}^{(t)} = \max_{k\in[t-W+1,\,t]} \chi^{(k)}, \tag{38}$$

and W is a sliding window size. As such, the inertia weight is updated as Eq. (39):

$$w^{(t)} = w_{\min} + (w_{\max} - w_{\min}) \times \mathcal{K}\big((\hat{\mathcal{F}}^{(t)} - \hat{\Delta}^{(t)}) \times \kappa\big), \tag{39}$$

with

$$\mathcal{K}(z) = \frac{1}{1+e^{-z}}.$$

Further, the parameter κ controls the sensitivity of the inertia update and initially $w^{(0)}$ is initialized as Eq. (40):

$$w^{(0)} = \frac{(w_{\max} + w_{\min})}{2}. \tag{40}$$

It is worth noting that:

- If $\hat{\mathcal{D}}_t \gg \hat{\Delta}_t$ (high diversity but weak improvement), the swarm explores without significant progress. In this case, the inertia weight decreases toward $w_{\min}$ to promote exploitation.
- If $\hat{\mathcal{D}}_t \ll \hat{\Delta}_t$ (low diversity but strong improvement), the swarm converges productively. The inertia weight is maintained at a moderate level to preserve stable convergence.
- If $\hat{\mathcal{D}}_t \approx \hat{\Delta}_t$, a balanced search state is achieved, leading to $w^{(t)} \approx (w_{\min} + w_{\max})/2$.
- If both $\hat{\mathcal{D}}_t$ and $\hat{\Delta}_t$ are small, stagnation is detected and stronger exploration mechanisms such as Lévy flight perturbation are triggered.

3) Lévy Flight Perturbation (LFP): When the swarm stagnates or loses diversity within τ_s consecutive iterations, randomness is injected using a Lévy flight to escape local minima. A β-index Lévy step is generated as [23] [see Eq. (41)]:

$$L_{j,\beta} = \frac{u_j}{|v_j|^{1/\beta}}, \quad j \le d^*, \tag{41}$$

where, u_j and v_j are two Gaussian distributions of zero mean [see Eq. (42)]:

$$u_j \sim \mathcal{N}(0, \sigma_u^2), \quad v_j \sim \mathcal{N}(0,1), \tag{42}$$

and

$$\sigma_u = \left[\frac{\Gamma(1+\beta)\sin(\pi\beta/2)}{\Gamma\big(\frac{1+\beta}{2}\big)\,\beta\,2^{(\beta-1)/2}}\right]^{1/\beta},$$

where, $\Gamma(\cdot)$ is the Gamma function. Using the Gamma–Legendre duplication formula [52] $\Gamma(1+\beta)/\Gamma\big(\frac{1+\beta}{2}\big) = \frac{2^\beta}{\sqrt{\pi}}\Gamma\big(\frac{\beta}{2}+1\big)$ along with the identity $\Gamma\big(\frac{\beta}{2}+1\big) = \frac{\beta}{2}\Gamma\big(\frac{\beta}{2}\big)$, σ_u can be reduced to Eq. (43):

$$\sigma_u = \sqrt{2}\left[\frac{\sin(\pi\beta/2)}{\sqrt{2\pi}}\right]^{1/\beta}. \tag{43}$$

The particle is then perturbed around the current global best [see Eq. (44)]:

$$\mathbf{p}_i^{(t)} \leftarrow \mathbf{p}_i^{(t)} + \alpha_L\,\mathbf{L}_\beta \odot \big(\mathbf{p}_i^{(t)} - \mathbf{g}^{(t)}\big), \tag{44}$$

where, $\mathbf{L}_\beta = (L_{1,\beta}, \ldots, L_{d^*,\beta})^T$, $\odot$ denotes element-wise (Hadamard) multiplication, and α_L is a significant small step size scaling factor. This introduces long-range jumps when the search becomes trapped.

4) Elite-Guided Reinitialization (EGR): To further prevent premature convergence, every $\mathcal{T}_{\text{EGR}}$ iterations, the worst-performing $N_{\textbf{elite}}$ particles (i.e., with highest fitness) are reinitialized around the global elite using Gaussian-based EGR [see Eq. (45)]:

$$\mathbf{p}_{\vartheta_i}^{(t)} = \mathbf{g}_{\text{best}} + \boldsymbol{\xi}, \quad 1 \leq i \leq N_{\textbf{elite}}, \tag{45}$$

where [see Eq. (46)],

$$\boldsymbol{\xi} \sim \mathcal{N}\big(\mathbf{0},\, \sigma_{\text{EGR}}^2 \mathbf{I}_{d^*}\big)\,, \tag{46}$$

$$f(\mathbf{p}_{\vartheta_1}^{(t)}) \geq f(\mathbf{p}_{\vartheta_2}^{(t)}) \geq \ldots \geq f(\mathbf{p}_{\vartheta_{\mathcal{S}}}^{(t)}), \vartheta_i \in \{1 \ldots \mathcal{S}\}, \tag{47}$$

and $\sigma_{\text{EGR}} = 0.1 \times \mathcal{F}^{(t)}$. This step increases population diversity by injecting new candidate solutions near the best-known region while still preserving exploitation [see Eq. (47)].

5) APSO velocity and position update: Leveraging the inertia weight assessed in Eq. (39), the velocity is updated as Eq. (48):

$$\begin{aligned}\mathbf{v}_i^{(t)} = {} & w^{(t-1)}\mathbf{v}_i^{(t-1)} + c_1\mathbf{r}_1 \odot \big(\mathbf{p}_{i,\text{best}}^{(t-1)} - \mathbf{p}_i^{(t-1)}\big) \\ & + c_2\mathbf{r}_2 \odot \big(\mathbf{g}_{\text{best}}^{(t-1)} - \mathbf{p}_i^{(t-1)}\big),\end{aligned} \tag{48}$$

where,

- $\mathbf{g}_{\text{best}}^{(\tau)} = \min_{1 \leq \tau \leq t} \mathbf{g}^{(t)}$ is the global best position until t.
- c_1 and c_2 are the cognitive and social acceleration coefficients, respectively;
- $\mathbf{r}_1, \mathbf{r}_2 \sim \mathcal{U}(0,1)^{d^\star}$ are independent random vectors introducing stochastic exploration.

To prevent the uncontrolled growth of the velocity and to stabilize the search process, a component-wise bound is imposed as:

$$v_{i,d}^{(t)} = \min\Big(\max\big(v_{i,d}^{(t)}, -v_{\max}\big), v_{\max}\Big), \quad 1 \leq d \leq d^\star. \tag{49}$$

As such, the particle position is updated as follows [see Eq. (50)]:

$$\mathbf{p}_i^{(t)} = \mathbf{p}_i^{(t-1)} + \mathbf{v}_i^{(t)}, 1 \leq i \leq \mathcal{S}. \tag{50}$$

Similarly to Eq. (49), the positions are projected onto the feasible search space using a component-wise clipping operator [see Eq. (51)]:

$$p_{i,d}^{(t)} = \min\Big(\max\big(p_{i,d}^{(t)}, -p_{\max}\big), p_{\max}\Big), \quad 1 \leq d \leq d^\star. \tag{51}$$

Algorithm 2 presents the proposed APSO framework for global optimization. It iteratively updates particle velocities and positions using adaptive inertia, stochastic perturbation, and elite-guided reinitialization to enhance exploration and prevent premature convergence. The algorithm returns the optimal global best solution $\mathbf{g}_{\text{best}}$, which initializes the subsequent fine-tuning stage.

Algorithm 2 Adaptive PSO

Input: $\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{val}}, p_{\max}, v_{\max}, \mathcal{S}, \mathcal{T}_{\text{APSO}}, d^*, c_1, c_2, \kappa, \tau_s$ $\mathcal{T}_{\text{SGD}}, \varepsilon, W, w_{\min}, w_{\max}, \beta, \alpha_N, \sigma_{\text{EGR}}, \mathcal{T}_{\text{EGR}}, N_{\textbf{elite}}$
Output: $\mathbf{g}_{\text{best}}$ //Global best position
for $i = 1$ to $\mathcal{S}$ **do** //Initialize swarm position and speed
  Generate randomly $\mathbf{p}_i^{(0)}, \mathbf{v}_i^{(0)}$ //using Eqs. (29)-(30)
  Initialize $\mathbf{p}_{i,\text{best}} \leftarrow \mathbf{p}_i^{(0)}$
  Evaluate the fitness $f(\mathbf{p}_i^{(0)})$
  Initialize $Stag_i \leftarrow 0$ //Stagnation counters $\leq \tau_s$
end for
Deduce $\mathbf{g}^{(0)}$ //using Eq. (32)
Initialize $\mathbf{g}_{\text{best}} \leftarrow \mathbf{g}^{(0)}$
Compute σ_u //using Eq. (43)
Initialize inertia weight $\omega^{(0)}$ //using Eq. (40)
for $t = 1$ to $\mathcal{T}_{\text{APSO}}$ **do**
  for $i = 1$ to $\mathcal{S}$ **do**
    Generate two d^*-vectors $\mathbf{r}_1$ and $\mathbf{r}_2$
    Update $\mathbf{v}_i^{(t)}$ //using Eq. (48) and then (49)
    Update $\mathbf{p}_i^{(t)} \leftarrow \mathbf{p}_i^{(t-1)} + \mathbf{v}_i^{(t)}$ //As in (50) then (51)
    if $f(\mathbf{p}_i^{(t)}) < f(\mathbf{p}_{i,\text{best}})$ **then**
      $\mathbf{p}_{i,\text{best}} \leftarrow \mathbf{p}_i^{(t)}$; $Stag_i \leftarrow 0$
    else
      $Stag_i \leftarrow Stag_i + 1$
      if $Stag_i \geq \tau_s$ **then** //Apply LFP
        for $j = 1$ to d^* **do**
          Generate u_j, v_j //using (42)
          Deduce $L_{j,\beta}$ //using (41)
        end for
        Update $\mathbf{p}_i^{(t)}$ //using (44)
        Re-evaluate $f(\mathbf{p}_i^{(t)})$ and re-calculate $\mathbf{p}_{i,\text{best}}$
        $Stag_i \leftarrow 0$
      end if
    end if
    if $f(\mathbf{p}_{i,\text{best}}) < f(\mathbf{g}_{\text{best}})$ **then** //Update $\mathbf{g}_{\text{best}}$: (32)
      $\mathbf{g}_{\text{best}} \leftarrow \mathbf{p}_{i,\text{best}}$;
    end if
  end for
  Compute $\Delta^{(t)}$ //using Eq. (34)
  Evaluate $\bar{\mathbf{p}}_{\text{best}}^{(t)}$ and Deduce $\mathcal{F}^{(t)}$ //using (36),(35)
  Compute $\Delta_{\min}^{(t)}, \Delta_{\max}^{(t)}, \mathcal{F}_{\min}^{(t)}, \mathcal{F}_{\max}^{(t)}$ //using Eq. (37)
  Deduce $\hat{\Delta}^{(t)}, \hat{\mathcal{F}}^{(t)}$ //using Eq. (38)
  Deduce $w^{(t)}$ //using Eq. (39)
  //Apply EGR
  Set $\sigma_{\text{EGR}} \leftarrow 0.1 \times \mathcal{F}^{(t)}$
  if $t \bmod \mathcal{T}_{\text{EGR}} = 0$ **then**
    Sort $\mathbf{p}_i^{(t)}$ in descending order of fitness //as in (47)
    for $i = 1$ to N_{elite} **do**
      Generate randomly $\boldsymbol{\xi}$ //as in (46)
      Update $\mathbf{p}_{\vartheta_i}^{(t)}$ //as in (45)
      if $f(\mathbf{p}_{\vartheta_i}^{(t)}) < f(\mathbf{p}_{\vartheta_i,\text{best}})$ **then**
        $\mathbf{p}_{\vartheta_i,\text{best}} \leftarrow \mathbf{p}_{\vartheta_i}^{(t)}$
      end if
    end for
  end if
  Recompute $\mathbf{g}_{best}$ from $\mathbf{p}_{\vartheta_i,\text{best}}$
  //Save position and velocity for subsequent iteration
  $\mathbf{p}_i^{(t+1)} \leftarrow \mathbf{p}_i^{(t)}, \quad \mathbf{v}_i^{(t+1)} \leftarrow \mathbf{v}_i^{(t)}$
end for
return $\mathbf{g}_{\text{best}}$

6) SGD-Based fine-tuning: After the APSO stage described in Section III-D, the optimal solution $\mathbf{g}_{\text{best}} \in \mathbb{R}^{d^*}$ is reshaped to initialize the parameters of a feedforward neural network with architecture $[d^*, n_1, 1]$, where d^* is the input dimension and n_1 denotes the number of hidden neurons. As such, each training sample is represented by a reduced feature vector $\mathbf{s} \in \mathbb{R}^{d^*}$ and $\mathbf{g}_{\text{best}}$ is partitioned into structured parameter blocks [see Eq. (52)]:

$$\mathbf{g}_{\text{best}} = [\mathbf{g}_{W_1}; \mathbf{g}_{b_1}; \mathbf{g}_{w_2}; g_{b_2}], \tag{52}$$

where, $\mathbf{g}_{W_1} \in \mathbb{R}^{n_1 \times d^*}$, $\mathbf{g}_{b_1} \in \mathbb{R}^{n_1}$, $\mathbf{g}_{w_2} \in \mathbb{R}^{n_1}$, and $g_{b_2} \in \mathbb{R}$.

The forward propagation step computes the hidden representation via a linear transformation followed by a ReLU activation. The output layer applies the sigmoid activation, defined in Section III-D2, to produce the final prediction [see Eq. (53)]:

$$\hat{y} = \mathcal{K}(\mathbf{g}_{w_2}^\top \mathbf{h} + g_{b_2}), \tag{53}$$

where [see Eq. (54)],

$$\mathbf{h} = \max(\mathbf{0}, \mathbf{g}_{W_1}\mathbf{s} + \mathbf{g}_{b_1}), \tag{54}$$

The complete trainable parameter vector is defined as Eq. (55):

$$\boldsymbol{\theta} = \{\mathbf{g}_{W_1}; \mathbf{g}_{b_1}; \mathbf{g}_{w_2}; g_{b_2}\}, \tag{55}$$

initialized at Eq. (56):

$$\boldsymbol{\theta}_0 = \text{reshape}(\mathbf{g}_{\text{best}}). \tag{56}$$

Obviously, $\mathbb{R}^{|\boldsymbol{\theta}|} = n_1 d^* + 2n_1 + 1$. The model is trained via backpropagation [46] by minimizing the binary cross-entropy loss with L_2 regularization:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{j=1}^{|\mathcal{D}_{\text{tr}}|} \underbrace{[-y_j \log \hat{y}_j - (1-y_j)\log(1-\hat{y}_j)]}_{\triangleq \mathcal{E}_j} - \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2, \tag{57}$$

where, $|\mathcal{D}_{\text{tr}}|$ is the number of training samples and λ controls the regularization strength.

At each iteration, the gradient of the loss function in Eq. (57) is computed over a mini-batch $\mathcal{B} \subset \mathcal{D}_{\text{tr}}$ of fixed size $|\mathcal{B}|$ [see Eq. (58)]:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \nabla_{\boldsymbol{\theta}} \mathcal{E}_j + \lambda \boldsymbol{\theta}, \tag{58}$$

where, ∇ denotes the gradient operator and $\mathcal{E}_j$ denotes the binary cross-entropy loss for sample j.

Optimization is performed using stochastic gradient descent [34], [35] with momentum [47]:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{u}_t, 0 \le t < \mathcal{T}_{\text{SGD}} - 1, \tag{59}$$

where, $\mathbf{u}_t \in \mathbb{R}^{|\boldsymbol{\theta}|}$, defined as:

$$\begin{cases} \mathbf{u}_t = \mu_m \mathbf{u}_{t-1} + \eta_\ell \nabla_{\boldsymbol{\theta}} \mathcal{L}, & 1 \le t \le \mathcal{T}_{\text{SGD}} \\ \mathbf{u}_0 = \mathbf{0} \end{cases}. \tag{60}$$

where, η_ℓ is the learning rate and μ_m is the momentum coefficient. Mini-batch training with a fixed batch size is used, and early stopping based on validation performance prevents overfitting.

Thus, the SGD stage iteratively applies the parameter update in Eq. (59) with momentum accumulation as defined in Eq. (60) to refine the APSO-initialized parameters and obtain the final optimized model.

Algorithm 3 summarizes the complete fine-tuning procedure.

Algorithm 3 SGD Fine-Tuning

Input: $\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{val}}, \mathbf{g}_{\text{best}}, \lambda, \eta_\ell, \mu_m, T_{\text{SGD}}, |\mathcal{B}|, T_{\text{pat}}$
Output: $\boldsymbol{\theta}^*$
$\boldsymbol{\theta}_0 \leftarrow \text{reshape}(\mathbf{g}_{\text{best}})$ //as in (55)
$\mathbf{u}_0 \leftarrow \mathbf{0}$
$\mathcal{L}^* \leftarrow \infty$, wait $\leftarrow 0$
for epoch $= 1$ to $\mathcal{T}_{\text{SGD}}$ **do**
  for each mini-batch $\mathcal{B}$ in epoch **do** // $\left\lfloor \frac{|\mathcal{T}_{\text{SGD}}|}{|\mathcal{B}|} \right\rfloor$ times
    $\mathbf{g}_t \leftarrow \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \nabla_{\boldsymbol{\theta}} \mathcal{L} + \lambda \boldsymbol{\theta}$
    $\mathbf{u}_t \leftarrow \mu_m \mathbf{u}_{t-1} + \eta_\ell \mathbf{g}_t$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \mathbf{v}_t$
  end for
  $\mathcal{L}_{\text{val}} \leftarrow \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_{\text{val}})$
  if $\mathcal{L}_{\text{val}} < \mathcal{L}^*$ **then**
    $\mathcal{L}^* \leftarrow \mathcal{L}_{\text{val}}$
    $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}$
    wait $\leftarrow 0$
  else
    wait $\leftarrow$ wait $+ 1$
    if wait $\ge T_{\text{pat}}$ **then**
      break
    end if
  end if
end for
return $\boldsymbol{\theta}^*$

IV. Experimental Setup and Data Description

This section describes the dataset, preprocessing pipeline, and experimental configuration used to evaluate the proposed framework. All preprocessing steps were performed using the training data only to avoid information leakage.

A. Dataset Description

The dataset contains $10,247$ recruitment records collected from a telecommunications company in North Africa between 2018 and 2022. After removing duplicates and incomplete evaluations (minimum six months of employment), the final dataset was retained for analysis.

The data includes 14 features, as outlined in Table I, grouped into four categories:

- Demographic (e.g., age, gender, marital status),
- Educational (e.g., education level, specialization, certifications),
- Professional (e.g., experience, salary, seniority, language proficiency),

TABLE I. DATASET DESCRIPTION GROUPED BY FEATURE CATEGORIES ($n = 10{,}247$)

Category	Feature	Value Range / Distribution
Demographic Features		
	Age	22–58 (mean: 31.4, SD: 6.2)
	Gender	Male (54%) / Female (46%)
	Marital Status	Single (61%) / Married (39%)
Educational Features		
	Education Level	Bachelor (58%), Master (35%), PhD (7%)
	Specialization	Tech (42%), Econ (31%), HR (15%), Law (12%)
	Certifications	Yes (34%) / No (66%)
Professional Features		
	Experience	0–15 years (mean: 4.8, SD: 3.1)
	Salary	4,000–18,000 MAD (mean: 8,200, SD: 2,800)
	Language Proficiency	1–5 (mean: 3.2, SD: 1.1)
Behavioral Features		
	Communication	1–5 (mean: 3.6, SD: 0.9)
	Motivation	Yes (73%) / No (27%)
	Remote Work	Yes (68%) / No (32%)
	Physical Ability	Yes (91%) / No (9%)
Target Variable		
	Performance	BA (38%) / Good (62%)

- Behavioral (e.g., communication skills, motivation, remote work adaptability, physical ability).

The target variable represents job performance and is defined as a binary label: *Below Average (BA)* and *Good*.

It should be noted that binary performance labels derived from HR evaluations are inherently subject to noise, as they reflect subjective managerial assessments that may vary across evaluators and organizational contexts. This label noise imposes an upper bound on achievable predictive accuracy and should be considered when interpreting the reported results. The 76.8% accuracy achieved by the proposed framework may partly reflect this noise ceiling rather than a fundamental model limitation.

Ethical approval was obtained from the institutional review board. All personal identifiers were removed, and data handling followed privacy regulations.

B. Data Preprocessing

The dataset was divided into 80% training ($n = 8{,}197$) and 20% test ($n = 2{,}050$) using stratified sampling. A 10-fold stratified cross-validation procedure was applied on the training set.

The preprocessing pipeline consists of the following steps:

- *Missing values:* Missing entries in several attributes are estimated using KNN imputation [37], [38] with k_{nn} neighbors. For a missing value $w_{i,j}$, the estimate is computed as Eq. (61):

$$\hat{x}_{i,j} = \frac{1}{k_{\text{nn}}} \sum_{\ell \in \mathcal{N}_{k_{\text{nn}}}(i)} x_{\ell,j}, \tag{61}$$

  where, $\mathcal{N}_{k_{\text{nn}}}(i)$ denotes the set of the k_{nn} nearest samples to observation i.

- *Encoding:* Binary variables are encoded as $\{0, 1\}$, while categorical attributes with multiple categories are transformed using one-hot encoding.

- *Scaling:* Continuous variables are normalized using Min–Max scaling [see Eq. (62)]:

$$x_{i,j}^{\text{scaled}} = \frac{x_{i,j} - x_j^{(\min)}}{x_j^{(\max)} - x_j^{(\min)}}. \tag{62}$$

- *Outlier handling:* Extreme values are capped using the interquartile range (IQR) rule [40], affecting approximately 1.4% of the observations.

All transformations were fitted on training data and applied to validation and test sets to prevent data leakage.

C. Baseline Models and Configurations

Classical machine learning models are used as benchmarks to evaluate the effectiveness of the proposed framework, including RF [18], SVM [49], and gradient boosting methods [26], [41]. The standard ANN baseline employs the Adam optimizer [36] with dropout regularization [45]. Xavier initialization [48] is used for SGD-only baselines. All implementations use scikit-learn [39] and custom Python modules. Rather than detailing their internal formulations, their feature usage, architectures, and optimization strategies are summarized in Table II.

TABLE II. MODEL CONFIGURATIONS USED FOR BENCHMARKING

Model	*Features*	*Architecture*	*Optimization*
RF	14	500 trees	–
XGBoost	14	500 trees	–
SVM	14	RBF kernel	SMO
Std. ANN	14	$(14 - 21 - 1)$	Adam
LR-ANN+SGD	LR (12)	$(12 - 18 - 1)$	SGD
LR-ANN+PSO-SGD	LR (12)	$(12 - 18 - 1)$	PSO $\rightarrow$ SGD
SW-MCFS-APSO-SGD	**SW-MCFS**	**Adaptive**	**APSO $\rightarrow$ SGD**

D. Experimental Setup

This subsection summarizes the experimental configuration used to evaluate the proposed framework. All hyperparameters for Algorithm 1 to Algorithm 3 are reported in Table III. These settings were fixed across all experiments to ensure fair comparison and reproducibility.

The evaluation protocol is based on:

- *Cross-validation:* 10-fold stratified cross-validation on the training set.
- *Performance metrics:* Accuracy, Precision, Recall, Specificity, and F1-Score.
- *Statistical testing:* McNemar's test for pairwise comparison on the held-out test set and the Friedman test with Nemenyi post-hoc analysis across folds. The significance level is set to $\alpha = 0.05$.
- *Ablation analysis:* Each ablation configuration is repeated 10 times with different random seeds to assess stability and robustness.

V. RESULTS

Comprehensive experiments are conducted to assess feature selection quality, predictive performance, statistical significance, robustness, interpretability, and computational efficiency of the proposed framework.

TABLE III. PARAMETER SETTINGS FOR ALGORITHM 1, 2, AND 3

Alg.	*Param.*	*Val.*	*Param.*	*Val.*	*Param.*	*Val.*
1	K	10	N	10247	d	14
	γ	5	η	0.05	R	30
2	$p_{\max}$	2	$v_{\max}$	1	$\mathcal{T}_A$	200
	$\mathcal{S}$	20	d^*	10	ϵ	10^{-10}
	W	5	κ	10	$w_{\min}$	0.2
	$w_{\max}$	0.9	β	1.5	τ_s	5
	α_L	0.01	N_e	$\lfloor \mathcal{S}/5 \rfloor$	$\mathcal{T}_E$	10
	c_1	2.0	c_2	2.0	η_ℓ	0.01
	μ_m	0.9	$\lvert\mathcal{B}\rvert$	32	T_{pat}	15

A. Feature Selection Analysis

1) Method stability and consensus weights: Table IV reports the cross-validation stability scores ρ_m and the corresponding consensus weights ω_m computed for each feature scoring method.

TABLE IV. STABILITY AND CONSENSUS WEIGHTS FOR SW-MCFS

Method	ρ_m	ω_m
Mutual Information	0.91	0.31
Wald Significance	0.87	0.24
Fisher Discriminant Loading	0.84	0.20
Permutation Importance	0.89	0.25

Mutual Information achieves the highest stability, resulting in the largest weight. In contrast, Fisher Discriminant Loading exhibits the lowest stability and is automatically down-weighted. This demonstrates that the proposed aggregation mechanism effectively prioritizes reliable scoring methods.

2) Feature relevance and selection outcome: Table V presents the normalized feature relevance scores and the aggregated score $\overline{S}_{\alpha_j,w}$.

TABLE V. REDUCED $d^* = 10$ FEATURES USING THE SW-MCFS ALGORITHM.

j	*Feature*	$\overline{S}^{(m,n)}_{1,\alpha_j}$	$\overline{S}^{(m,n)}_{2,\alpha_j}$	$\overline{S}^{(m,n)}_{3,\alpha_j}$	$\overline{S}^{(m,n)}_{4,\alpha_j}$	$\overline{S}_{\alpha_j,w}$
1	*Communication*	1.00	0.92	0.88	1.00	**0.96**
2	*Experience*	0.87	1.00	0.91	0.82	**0.90**
3	*Seniority*	0.82	0.85	1.00	0.78	**0.86**
4	*Salary*	0.79	0.78	0.82	0.85	**0.81**
5	*Education*	0.71	0.74	0.69	0.73	**0.72**
6	*Language Prof.*	0.68	0.65	0.61	0.71	**0.67**
7	*Specialization*	0.62	0.58	0.55	0.64	**0.60**
8	*Remote Work*	0.55	0.52	0.48	0.58	**0.54**
9	*Certifications*	0.48	0.51	0.44	0.46	**0.47**
10	*Age*	0.42	0.45	0.38	0.40	**0.41**
11	*Gender*	0.28	0.31	0.22	0.25	0.27
12	*Marital Status*	0.24	0.27	0.19	0.21	0.23
13	*Motivation*	0.18	0.15	0.12	0.20	0.16
14	*Physical Ability*	0.12	0.08	0.14	0.11	0.11

Using the adaptive elbow threshold defined in Eq. (27) with $\tau^* = 0.34$, the algorithm selects $d^* = 10$ features. Sensitive demographic attributes such as Gender and Marital Status are automatically discarded due to their low aggregated relevance.

Compared with baseline selection methods, SW-MCFS preserves informative predictors (e.g., Experience, Education, Communication) while filtering low-signal attributes. This improves both interpretability and robustness against potential bias.

TABLE VI. PERFORMANCE COMPARISON (10-FOLD CV, $n = 8{,}197$) MEAN $\pm$ SD

Model	*Acc.*	*Prec.*	*Rec.*	*Spec.*	*F1*
Random Forest	72.1 ± 1.8	70.3 ± 2.1	73.5 ± 2.4	69.8 ± 2.0	71.8 ± 1.9
XGBoost	73.8 ± 1.5	72.1 ± 1.9	75.2 ± 2.0	71.4 ± 1.7	73.6 ± 1.6
SVM	70.4 ± 2.3	68.9 ± 2.5	71.2 ± 2.8	69.1 ± 2.4	70.0 ± 2.3
Standard ANN	73.2 ± 1.9	71.5 ± 2.2	74.6 ± 2.1	70.9 ± 2.0	73.0 ± 1.9
LR-ANN + SGD	74.6 ± 1.6	73.1 ± 1.8	75.6 ± 1.9	72.8 ± 1.7	74.3 ± 1.6
LR-ANN + PSO-SGD	75.2 ± 1.4	73.6 ± 1.7	76.3 ± 1.8	73.5 ± 1.6	74.9 ± 1.5
SW-MCFS-ANN + SGD	76.2 ± 1.3	74.8 ± 1.6	77.1 ± 1.7	74.5 ± 1.5	75.9 ± 1.4
SW-MCFS-APSO-SGD	$\mathbf{76.8 \pm 1.2}$	$\mathbf{75.3 \pm 1.5}$	$\mathbf{77.8 \pm 1.6}$	$\mathbf{75.2 \pm 1.4}$	$\mathbf{76.5 \pm 1.3}$

B. Model Comparison Results

Table VI compares the predictive performance of classical machine learning models and neural networks under different optimization strategies. Performance is evaluated using Accuracy, Precision, Recall, Specificity, and F1-score. Results are reported as mean $\pm$ standard deviation (SD) over 10-fold stratified cross-validation.

The proposed SW-MCFS-APSO-SGD framework achieves the highest accuracy while maintaining the lowest variance across folds. The reduced standard deviation indicates stable convergence and strong generalization ability.

C. Statistical Significance Analysis

To verify whether performance improvements are statistically significant, pairwise and global non-parametric tests are performed. Table VII reports the McNemar's test results comparing the proposed SW-MCFS-APSO-SGD against all competing models.

TABLE VII. MCNEMAR'S TEST COMPARING SW-MCFS-APSO-SGD AGAINST COMPETING MODELS ($p < 0.05$).

Comparison (vs.)	χ^2	*p-value*
Random Forest	16.42	< 0.001
XGBoost	6.31	0.012
SVM	22.87	< 0.001
Standard ANN	8.14	0.004
LR-ANN + SGD	4.85	0.028
LR-ANN + PSO-SGD	4.18	0.041
SW-MCFS-ANN + SGD	3.92	0.048

The Friedman test yields $\chi^2(7) = 45.23$ with $p < 0.001$, confirming overall statistical differences among models. Post-hoc Nemenyi analysis further ranks the proposed framework as the best-performing method with a critical difference of 1.42 at $\alpha = 0.05$.

These results confirm that the improvements of the proposed approach are statistically significant.

D. Ablation Studies

Ablation experiments are conducted to quantify the contribution of individual components.

1) Ablation study for feature selection methods: Table VIII reports the performance of different feature selection strategies, highlighting the contribution of the proposed SW-MCFS method compared to individual and baseline selection techniques.

TABLE VIII. ABLATION A1: FEATURE SELECTION COMPARISON (ALL WITH SGD OPTIMIZER).

Selection Method	*# feat.*	*Accuracy*	*F1*
MI only	11	74.8 ± 1.5	74.5
Wald (LR) only	12	74.6 ± 1.6	74.3
FDL (DA) only	8	68.5 ± 2.1	67.7
PI only	11	75.0 ± 1.5	74.7
Equal-weight consensus	10	75.8 ± 1.4	75.5
SW-MCFS (ours)	**10**	$\mathbf{76.2 \pm 1.3}$	**75.9**

SW-MCFS outperforms single-method selection and equal-weight aggregation. The stability-based weighting contributes to consistent improvements in predictive accuracy.

2) Optimization ablation: Table IX presents the contribution of each component in the proposed framework, evaluating the performance degradation when individual modules are removed.

TABLE IX. ABLATION A2: OPTIMIZER COMPARISON (ALL WITH SW-MCFS FEATURES).

Optimizer	*Accuracy*	*Epochs to 95%*	*Time (s)*
SGD only (Xavier)	76.2 ± 0.8	50.2 ± 4.3	138 ± 8
PSO $\rightarrow$ SGD	76.6 ± 0.6	38.1 ± 3.7	187 ± 12
APSO $\rightarrow$ SGD	$\mathbf{76.8 \pm 0.5}$	$\mathbf{34.5 \pm 3.1}$	$\mathbf{195 \pm 14}$

APSO improves convergence speed and final accuracy compared with standard PSO and SGD-only training with Xavier initialization [48]. The hybrid strategy achieves better optimization stability.

3) Component contribution: Table X quantifies the effect of removing individual components from the proposed framework to assess their relative contribution to the overall performance.

TABLE X. CONTRIBUTION OF INDIVIDUAL COMPONENTS

Configuration	*Accuracy*	Δ *vs. full*
LR + SGD (baseline)	74.6	-2.2
LR + PSO-SGD	75.2	-1.6
SW-MCFS + SGD	76.2	-0.6
SW-MCFS + PSO-SGD	76.6	-0.2
SW-MCFS + APSO-SGD (full)	**76.8**	—

Both feature selection and optimization contribute positively and additively to the overall performance improvement.

E. SHAP Feature Importance

Table XI reports the global SHAP importance scores [50] computed on the trained SW-MCFS-APSO-SGD model, ranking features according to their mean absolute contribution.

TABLE XI. GLOBAL SHAP IMPORTANCE FOR THE PROPOSED MODEL. DEMOGRAPHIC ATTRIBUTES SUCH AS GENDER AND MARITAL STATUS ARE EXCLUDED BY SW-MCFS.

Feature	*Mean* $\lvert\phi_i\rvert$	*Rank*
Communication Skills	0.358	1
Experience	0.301	2
Seniority Level	0.284	3
Salary	0.259	4
Education Level	0.215	5
Language Proficiency	0.194	6
Specialization	0.172	7
Remote Work	0.148	8
Certifications	0.078	9
Age	0.065	10

The results confirm that SW-MCFS automatically excludes sensitive demographic attributes, while communication skills, experience, and seniority remain dominant predictors.

F. Held-Out Test Set Performance

Table XII compares the proposed framework against baseline models on the independent held-out test set.

TABLE XII. PERFORMANCE ON THE INDEPENDENT TEST SET ($n = 2{,}050$)

Model	*Acc.*	*Prec.*	*Rec.*	*F1*
Random Forest	71.8	69.9	73.2	71.5
XGBoost	73.5	71.8	74.9	73.3
SVM	70.1	68.5	70.8	69.6
Standard ANN	72.9	71.2	74.3	72.7
LR-ANN + PSO-SGD	75.0	73.4	76.0	74.7
SW-MCFS-APSO-SGD	**76.5**	**75.1**	**77.6**	**76.3**

The proposed method achieves the highest performance across all metrics, demonstrating strong generalization capability.

G. Computational Efficiency

Table XIII reports the training and inference time of the proposed framework compared to competing models.

TABLE XIII. TRAINING INCLUDES SW-MCFS COMPUTATION ($\sim$24S). IDENTICAL INFERENCE FOR ALL ANNS.

Model	*Train (s)*	*Inference (s/1k)*
XGBoost	68	0.08
LR-ANN + PSO-SGD	187	0.05
SW-MCFS + SGD	162	0.05
SW-MCFS-APSO-SGD	**219**	**0.05**

Although the proposed model requires slightly higher training time due to APSO optimization, inference complexity remains unchanged while achieving superior performance.

VI. DISCUSSION

A. Principal Findings

This study highlights four key findings:

1) SW-MCFS significantly improves feature selection: The stability-weighted multi-criteria approach improves accuracy by 1.6% over single-method LR selection (Table X). Its effectiveness stems from aggregating heterogeneous relevance criteria, applying stability-based weighting to suppress unreliable scorers, and automatically determining the feature dimension via an adaptive threshold. It is important to emphasize that the novelty lies not in the individual scoring methods, which are well-established, but in their principled integration through stability-weighted consensus—a combination that produces emergent benefits including automatic dimensionality determination and demographic attribute filtering that no single method achieves independently.

2) APSO enhances optimization efficiency: The adaptive inertia strategy improves the exploration–exploitation balance, yielding 0.6% gain over SGD-only and 0.2% over standard PSO, with 31% faster convergence. Lévy-flight perturbations further improve robustness by enabling escape from local optima.

3) Observed exclusion of demographic attributes: SW-MCFS systematically excludes Gender and Marital Status, producing a feature set focused on job-relevant variables. This behavior emerges from consensus scoring rather than explicit constraints, as unstable and low-importance features are downweighted across folds. However, this observation should not be interpreted as a formal fairness guarantee; formal fairness metrics such as demographic parity and equalized odds were not computed in this study and remain a necessary direction for future validation [31].

4) Competitive performance against ensemble models: The proposed SW-MCFS–APSO–SGD framework achieves 76.8% accuracy, significantly outperforming XGBoost (73.8%, $p = 0.012$), demonstrating that well-designed neural pipelines can surpass strong ensemble baselines on structured HR data. It is worth noting that in predictive hiring, where target labels are derived from subjective performance evaluations, the inherent label noise imposes an accuracy ceiling that limits all models. Within this context, a 3% absolute improvement over XGBoost represents a meaningful and statistically significant gain [43].

B. Novelty Positioning

Table XIV contrasts the proposed framework with representative prior works. Existing studies typically address either multi-criteria feature selection or adaptive optimization separately. In contrast, the proposed framework jointly integrates stability-weighted multi-method feature selection with adaptive particle-based optimization in a unified learning pipeline.

TABLE XIV. POSITIONING OF THE PROPOSED FRAMEWORK RELATIVE TO PRIOR WORK.

Study	*Multi-FS*	*Adaptive PSO*	*HR Data*
Saeys et al. [20]	✓	–	–
Zhan et al. [22]	–	✓	–
Qin et al. [24]	–	–	✓
Wang et al. [25]	–	–	✓
Ours	✓	✓	✓

C. Limitations

Despite strong empirical performance, several limitations should be acknowledged.

1) Dataset scope: All experiments were conducted on a single-organizational dataset from a North African telecommunications company, which limits external generalizability. The abstract and claims throughout the study are qualified accordingly. Validation across multiple industries, geographic regions, and organizational cultures is required before broader applicability can be established.

2) Label quality: The binary performance labels are derived from subjective managerial evaluations, which constitute a significant source of noise in predictive hiring datasets. Inter-rater reliability and evaluation criteria consistency were not assessed. This noise ceiling contextualizes the 76.8% accuracy and suggests that absolute performance gains should be interpreted relative to this inherent limitation.

3) Fairness evaluation: Although demographic features (Gender, Marital Status) were automatically excluded by SW-MCFS due to low consensus scores, this observation does not constitute a formal fairness guarantee. No formal fairness metrics such as demographic parity, equalized odds, or disparate impact ratios were computed [31]. The exclusion of protected attributes from the feature set does not preclude indirect discrimination through proxy variables (e.g., salary or seniority may correlate with demographic characteristics). Formal fairness auditing remains a critical direction for deployment readiness.

4) Computational cost: SW-MCFS requires multiple feature scorers over K folds, and APSO introduces additional optimization overhead. Scalability to very large datasets (e.g., $> 100{,}000$ records) remains to be evaluated.

5) Hyperparameter sensitivity: Parameters such as γ, APSO settings, and population size were tuned empirically. A systematic sensitivity analysis would further strengthen robustness claims.

D. Ethical Considerations

The deployment of predictive hiring systems raises significant ethical concerns that extend beyond predictive accuracy. Algorithmic hiring tools are increasingly subject to regulatory scrutiny, including the European Union Artificial Intelligence Act [53], which classifies employment-related AI systems as high-risk and mandates transparency, human oversight, and bias auditing requirements [43].

Several ethical dimensions are relevant to the proposed framework. First, although SW-MCFS excludes Gender and Marital Status from the selected feature set, indirect discrimination may persist through correlated proxy variables. For instance, salary levels and seniority may encode historical gender-based disparities, enabling the model to implicitly learn demographic associations. Formal fairness auditing using established metrics (demographic parity, equalized odds, and disparate impact analysis) is essential before any operational deployment.

Second, the use of subjective performance labels as ground truth raises concerns about perpetuating existing organizational

biases. If historical evaluations systematically disadvantaged certain groups, the trained model may reproduce these patterns. Bias-aware label correction and calibration techniques should be explored in future work.

Third, transparency and explainability are critical for stakeholder trust. The SHAP-based interpretability analysis presented in this study provides a foundation for explaining individual predictions to candidates and hiring managers. However, the right to explanation and the principle of human-in-the-loop decision-making require that predictive scores serve as decision-support tools rather than automated gatekeepers.

E. Future Directions

Several extensions can enhance the proposed framework:

1) Fairness-aware extension: Incorporating an explicit fairness-aware scoring term into SW-MCFS would allow penalizing features strongly correlated with protected attributes [see Eq. (63)].

$$\overline{S}^{\text{fair}}_{\alpha_j,w} = \sum_m \omega_m \bar{S}^{(m)}_{\alpha_j} - \lambda_f \mathcal{F}_j, \tag{63}$$

where, λ_f controls the fairness–accuracy trade-off.

2) Deep architectures: Applying the SW-MCFS–APSO pipeline to deeper tabular models (e.g., TabNet [51] or transformer-based architectures) would test scalability to higher-dimensional parameter spaces.

3) Cross-domain validation: Benchmarking on public HR and structured prediction datasets (credit risk, medical classification, etc.) would further validate generalization capability.

4) Online feature selection: Developing an incremental version of SW-MCFS would enable dynamic feature relevance updates as new data becomes available.

VII. Conclusion

This study presents two complementary methodological contributions for predictive hiring: *Stability-Weighted Multi-Criteria Feature Selection (SW-MCFS)* and *Adaptive Particle Swarm Optimization (APSO)*. SW-MCFS aggregates heterogeneous feature scoring methods through stability-weighted consensus across cross-validation folds, producing robust feature subsets while automatically filtering low-signal and demographic attributes. APSO enhances neural network initialization by combining adaptive inertia control with Lévy-flight perturbations, enabling effective exploration of complex optimization landscapes.

The integrated SW-MCFS-APSO-SGD framework achieves 76.8% accuracy on 10,247 recruitment records, significantly outperforming XGBoost (73.8%, $p = 0.012$), standard PSO-SGD (75.2%, $p = 0.041$), and individual feature selection strategies. Ablation analyses confirm that SW-MCFS provides the largest gain ($+1.6\%$), while APSO further improves optimization performance ($+0.6\%$) and accelerates convergence by approximately 31%. These results should be interpreted in the context of inherent label noise in subjective performance evaluations, which imposes an accuracy ceiling on all models.

Notably, SW-MCFS excludes Gender and Marital Status through consensus scoring without explicit fairness constraints. While this observation is encouraging, it does not constitute a formal fairness guarantee, and formal fairness auditing remains necessary for deployment readiness. The SHAP-based interpretability analysis, which highlights Communication Skills, Experience, and Seniority as dominant predictors, provides a foundation for transparent decision-support in regulated HR applications.

The proposed framework is applicable to other classification tasks requiring robust feature selection and efficient neural optimization. Future work will extend the approach to multi-domain validation across diverse industries and geographic contexts, incorporate explicit fairness objectives with formal bias metrics, and explore scalability to deeper architectures and multimodal data.

Data Availability Statement

The dataset is derived from anonymized recruitment data collected at ENSIAS, Université Mohammed V – Rabat. Due to confidentiality agreements, the data cannot be publicly shared. Researchers may contact the corresponding author. Analysis scripts and SW-MCFS implementation are available upon reasonable request.

Author Contributions

Y.T.K. conceptualized the study, designed SW-MCFS and APSO, performed analysis, and drafted the manuscript. S.A. supervised, validated results, and revised the manuscript. Both authors approved the final version.

Funding

This research received no external funding.

Institutional Review Board Statement

Approved by the participating organization's internal review board (Protocol: HR-2023-015, 15 January 2023).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgment

The authors thank the HR department of the participating organization and the anonymous reviewers for constructive feedback.

References

[1] Z.-H. Zhou, K. Chen, and X. Y. Dai, "Data-driven intelligence in hiring and employee selection," *Int. J. Data Sci. Anal.*, vol. 11, no. 3, pp. 221–234, 2021.

[2] R. Chamorro-Premuzic, D. Winsborough, R. A. Sherman, and R. Hogan, "New talent signals: Shiny new objects or a brave new world?" *Ind. Organ. Psychol.*, vol. 9, no. 3, pp. 621–640, 2016.

[3] S. Pessach, G. Singer, D. Avrahami, H. C. Ben-Gal, E. Shmueli, and I. Ben-Gal, "Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming," *Decis. Support Syst.*, vol. 134, p. 113290, 2020.

[4] V. Jagwani, S. Meghani, K. Pai, and S. Dhage, "Resume evaluation through LDA and NLP for effective candidate selection," *arXiv:2307.15752*, 2023.

[5] E. Faliagka, A. Tsakalidis, and G. Tzimas, "An integrated e-recruitment system for automated personality mining and applicant ranking," *Internet Res.*, vol. 22, no. 5, pp. 551–568, 2012.

[6] P. Levy, O. Golan, and A. Schuster, "Enhancing interview evaluations with nonverbal behavior analysis," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 205–216, 2018.

[7] S. Strohmeier and F. Piazza, "Artificial intelligence techniques in human resource management—A conceptual exploration," in *Intelligent Techniques in Engineering Management*. Cham: Springer, 2015, pp. 149–172.

[8] J. Pfeffer and R. I. Sutton, "Evidence-based management," *Harvard Bus. Rev.*, vol. 84, no. 1, pp. 62–74, 2006.

[9] IBM Smarter Workforce Institute, *The Business Case for AI in HR*. Armonk, NY: IBM Corp., 2017.

[10] Y. Khallouk Temsamani and S. Achchab, "Artificial intelligence use in human resources management: Strategy and operation's impact," ENSIAS, Mohammed V University, Rabat, Morocco, 2021.

[11] N. Otani *et al.*, "Natural language processing for human resources," in *Proc. NAACL Industry Track*, 2025.

[12] K. Khelkhal and D. Lanasri, "Smart-Hiring: An explainable NLP pipeline for CV information extraction and job matching," *arXiv:2511.02537*, 2025.

[13] C. Gan *et al.*, "Application of large language models in resume screening and recruitment," *arXiv:2401.08315*, 2024.

[14] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in ML-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023.

[15] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Netw.*, vol. 4, 1995, pp. 1942–1948.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York: Springer, 2009.

[17] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ: Wiley, 2013.

[18] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[19] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[20] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Proc. ECML PKDD*, 2008, pp. 313–325.

[21] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *Proc. IEEE World Congr. Comput. Intell.*, 1998, pp. 69–73.

[22] Z.-H. Zhan, J. Zhang, Y. Li, and H. S.-H. Chung, "Adaptive particle swarm optimization," *IEEE Trans. Syst. Man Cybern. B*, vol. 39, no. 6, pp. 1362–1381, 2009.

[23] A. M. Jensi and S. Jeyabharathi, "Enhanced particle swarm optimization with Lévy flight for global optimization," *Appl. Soft Comput.*, vol. 43, pp. 248–261, 2016.

[24] C. Qin *et al.*, "Enhancing person-job fit for talent recruitment: An ability-aware neural network approach," *arXiv:1812.08947*, 2018.

[25] Z. Wang, W. Wei, C. Xu, J. Xu, and X.-L. Mao, "Person-job fit estimation with co-attention neural networks," *arXiv:2206.09116*, 2022.

[26] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.

[28] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 6, p. 066138, 2004.

[29] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[30] D. He, X. Yuan, and H. Li, "AI and talent acquisition: The emerging frontier," *J. Bus. Res.*, vol. 112, pp. 140–148, 2020.

[31] H. Heidari, A. Ferrario, and M. B. Zafar, "Fairness in machine learning: From statistical to causal definitions," *Data Min. Knowl. Discov.*, vol. 34, no. 2, pp. 453–495, 2020.

[32] T. Zimmermann, L. Kotschenreuther, and K. Schmidt, "Data-driven HR – Résumé analysis based on NLP and machine learning," *arXiv:1606.05611*, 2016.

[33] T. T. Nguyen and T. H. Cao, "Job prediction: From deep neural network models to applications," *arXiv:1912.12214*, 2019.

[34] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*, 2nd ed. Berlin: Springer, 2012, pp. 421–436.

[35] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv:1609.04747*, 2016.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[37] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, suppl. 1, pp. S412–S419, 2001.

[38] G. E. Batista and M. C. Monard, "A study of K-nearest neighbour as an imputation method," in *Proc. Brazilian Symp. Intell. Syst.*, 2003, pp. 251–260.

[39] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[40] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

[41] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 6638–6648.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[43] A. Raghavan, S. Mohan, and P. Stone, "Bias in automated hiring systems: A systematic review," *AI Ethics*, vol. 2, no. 3, pp. 421–439, 2022.

[44] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[46] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[47] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. Math. Phys.*, vol. 4, no. 5, pp. 1–17, 1964.

[48] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.

[49] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[50] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4765–4774.

[51] S. Ö. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proc. AAAI*, vol. 35, no. 8, 2021, pp. 6679–6687.

[52] R. N. Mantegna, "Fast, accurate algorithm for numerical simulation of Lévy stable stochastic processes," *Phys. Rev. E*, vol. 49, no. 5, pp. 4677–4683, 1994.

[53] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Off. J. Eur. Union*, L series, 2024.

GTMedoids: A New Grey Sheep Users Detection Approach

Bouchra Boualaoui[1], Ahmed Zellou[2], Lahbib Ajallouda[3]
SPM Research Team, ENSIAS, Mohammed V University in Rabat, Rabat, Morocco[1,2]
Poly-Disciplinary Faculty of Es-Semara, Ibn Zohr University Agadir, Morocco[3]

***Abstract*—Recommender systems have been developed to serve users and provide them with the best suggestions. Despite their success, offering fully identical recommendations to users' preferences remains a difficult task, where the complexity of human taste results in different challenges. Grey sheep user phenomena continues to be one of the most common, where the user is defined by his unique interactions with the system, making it difficult to associate with similar users, as he rarely agrees with them. In this study, we presented a new approach for identifying grey sheep users. It is based on the taste context and nature of user interaction with the system. We grouped similar users using an enhanced Kmedoids clustering method with a new dissimilarity metric and introduced a novel process to distinguish between users. The differentiation is achieved by assigning weights to each cluster based on how much it reflects the grey sheep user characteristics. We evaluated the efficiency of Grey Threshold Medoids (GTMedoids) using the FilmTrust and MovieLens 100k datasets. The results show the superior performance of our approach in detecting grey sheep users.**

***Keywords*—Recommender system; grey sheep users detection; clustering; Kmeans; Kmedoids; dissimilarity metric**

I. Introduction

Recommender systems (RS) have demonstrated their ability to retrieve what could be interesting to users and help improve the user experience in a digital world overwhelmed by the explosion of data and the speed with which data is growing. However, this efficiency is exposed to high degradation due to multiple challenges, related to sparsity [1], [2], scalability [3], or fairness and user satisfaction, of which we can cite grey sheep users (GSUs) [4], where the first step to accurately analyze their preferences is to detect them. Although grey sheep user detection is important, it has been neglected in the literature. In this regard, our motivation behind this work is the importance of fairness and equality treatment that should be attributed to recommender systems to always provide all users with accurate suggestions and not only serve part of them and discriminate against the rest. To accurately detect grey sheep users in a system, the following research questions have been formulated:

- RQ1: How can user similarity be redefined to fully capture user taste behavior?
- RQ2: Can grey sheep users be detected dynamically and effectively from the entire user population without predefined groups?
- RQ3: How does the presence of grey sheep users in a system affect recommendation accuracy?

In this study, we present GTMedoids as a new proposed solution for detecting grey sheep users in a recommender system. We clustered candidate grey sheep users based on their interactions with the system and considered the bi-polarity of human taste. Then, attribute to each cluster a weight, where only clusters with power weights less than a predefined threshold are finally detected as grey sheep clusters as they meet the criteria of this category of users. To prove the validity of our model, we ran a set of experiments on two different benchmarks and demonstrated the efficiency of our model in detecting grey sheep users.

The study is structured into several sections: Section II provides background information and reviews existing literature on grey sheep user detection. Section III details the proposed approach, GTMedoids, and examines its various components. Section IV analyzes the results obtained from the study. Finally, Section V concludes the work and outlines potential future research directions.

II. Related Work

The evolution of recommender systems can be marked with three generations, where the first generation was initiated to focus on content and collaborative techniques [5], [6], the second enhanced quality and precision by involving more contextual information, such as location and time [7], [8], while the third generation is known for leveraging the hidden semantics in supplementary sources of knowledge, including social networks and knowledge graphs [9]. However, during the course of this enhancement, major challenges were addressed. Yet, some of these challenges were less treated than others, regardless of their importance, especially grey sheep users.

Grey sheep users are defined as users with significant deviations from recognized preference patterns [10]. Due to their unique taste, they end up receiving inaccurate suggestions that do not match their interaction with the system [4]. Although the importance of the topic, as it has a direct impact on the system's accuracy and the quality of recommendations for all users, it is considered one of the least treated RS challenges in the literature.

To address the challenge of GSUs, detecting these users remains a crucial step. The detection techniques in the literature can be grouped into two main categories. Outlier detection and clustering techniques. The first group treats GSUs as anomalies and outliers due to their odd choices. These techniques rely on ratings distribution and user profiles [11], [12], [14]. However, other outlier techniques can be explored, as shared in [10], such

as angle-based outlier detection (ABOD) [15]. Regardless of the popularity of outlier detection methods as a way to identify grey sheep users in a system, their nature and the way that they are designed highly put the results under investigation, as the final detected set of users can also include those suffering from sparsity and sparsity issues due to their low density, while a grey sheep user is a user that is committed to the system but has different preferences [18], [11]. Adding to that, any data points that are close to a grey sheep user run the risk of being flagged as a grey sheep user [20]. And it is important to also mention that these methods need to treat the whole datasets, which can make the distance or density scores attached to all type of users even less meaningful [29].

On the other hand, the second category uses clustering approaches that are based on computing user similarities [16], [17], [19], [21], [22]. The primary limitation of these approaches is the fact that they cluster all the users of the system and view them all as candidates, which can lead to high computational costs and group all the final grey sheep users in one $(K+1)^{th}$ cluster, leading to a lack of clarity about the degree of grayness of the users, especially in the case where the system aims to remove them [14]. Moreover, all these works, regardless of their category, have neglected the polarity of human taste. However, it deserves attention that [4] highlighted the possibility of identifying GSUs using personality parameters but without real application, as they have used only the cosine similarity with KNN [23]. Therefore, we present GTMedoids, a clustering method that takes into account both sides of user feedback and also weights the degree of users' grayness.

III. GTMedoids Approach

In this section, we present our proposed GTMedoids approach and its different components. Fig. 1 provides a high-level overview of the approach and the interaction between its parts.

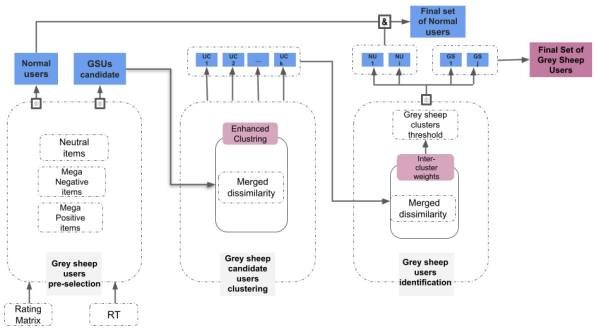


Fig. 1. Grey threshold medoids approach.

A. Grey Sheep Users Pre-Selection

Grey sheep users have unique and distinguished preferences, making the recommendation process more difficult and not as accurate as for normal users, who already have close users in terms of taste and similarities. Consequently, identifying these users should be based mainly on user preferences shared in the rating matrix [24]. Most studies measure preference based on the interaction with the item, where the existence of the interaction is interpreted as an interest and preference, ignoring the fact that an interaction with bad feedback should not be treated as a good one, as human taste is not only reflected in what he likes l but also in what he does not like l, which can be denoted as follows [see Eq. (1)]:

$$taste = \{d/d \in Detest\} \cup \{l/l \in Like\} \tag{1}$$

Taking this into consideration, grey sheep user identification should be based on both types of feedback; as a result, a GSU is then a user who highly disagrees with his community.

Clustering all users with prior knowledge that not all of them are GSUs is not efficient, as only concentrating on a set of users that can meet GSU characteristics; in particular that Kmeans variations are sensitive to outliers because of the mean computations [25]. In our case, outliers are normal users who more often share the same taste as the majority of the community. However, before deciding to exclude these users, we need to quantify the community's taste for comparison with individual preferences. We consider the global classification of items as collaborative signals contributed by all users in the system. An item can then be:

- Mega Positive: is an item that the majority has appreciated by giving it individual positive feedback Pos.
- Mega Negative: is an item that the majority has unappreciated by giving it an individual negative feedback Neg.
- Neutral: is an item having the same amount of individual positive and negative feedback.

The differentiation between positive and negative individual feedback is done based on a rating threshold RT as our approach treats the user-item rating matrix R [see Eq. (2)]:

$$feedback_{u,i} = \begin{cases} Pos & \text{if } r_{u,i} \geq RT \\ Neg & \text{otherwise} \end{cases} \tag{2}$$

where, $r_{u,i}$ is the rating given by the user u to the item i.

After computing the collaborative classification of each item, we refine our initial set to cluster by excluding normal users.

- Normal user: Frequently aligns with global preferences, showing a level of agreement that meets or exceeds the average.
- Grey sheep candidate user: Rarely aligned with the collaborative signals, meaning having an amount of less than the average of agreements.

The filtered users show potential to be grey sheep users, requiring further refinement as the actual feedback and ratings were not used to determine their greyness level or similarities with others. The subsequent step focuses on calculating user similarities and identifying their nearest neighbors.

B. Grey Sheep Candidate Users Clustering

To enhance the accuracy of user clustering, it is essential to include all rated items, even neutral ones. The proposed clustering approach modifies the Kmeans algorithm by applying a new dissimilarity metric and using actual user profiles as the centers of clusters instead of means and leads to more significant associations among user groups [31].

1) The merged dissimilarity metric: We have developed our own measurement metric for two reasons:

a) The multi-dimensional nature of human taste: The dissimilarity function must consider both aspects of taste, with the weight assigned to each depending on user preference. Some users prefer to share positive feedback, while others are more inclined to express negative feedback.

b) Reduce the unexplored items effect: One major issue with distance functions is their inability to differentiate items rated by only one user, leading to the same distance despite differing preferences. To address this, a contextual approach was adopted, where items are considered closer only if they share a collaborative taste. This involves incorporating feedback nature as context with weights, so items can be treated similarly for users if their ratings are closely aligned and they share similar feedback contexts. Fig. 2 illustrates how adding multi-dimensional context improves accuracy in understanding user preferences, as the Euclidean distance can misleadingly suggest similarity between items that do not truly reflect preferences, impacting the mining of user data, where this distance may treat items i_0 and i_1 as equivalent, while they do not truly reflect aligned user preferences. Consequently, this miscalibration leads to false similarities among users, such as u_0 and u_1.

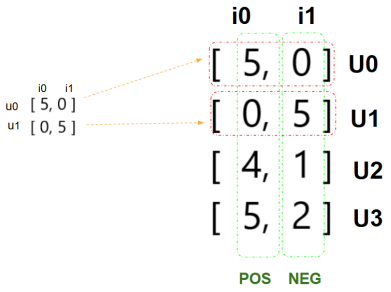


Fig. 2. The impact of the unexplored items on the dissimilarity measurement.

The final metric is a user-personalized metric that merges both sides of human tastes and can be presented as follows [see Eq. (3)]:

$$MergedD(u_a, u_b) = weight_{dislike} \times DisLikedD(u_a, u_b) + weight_{like} \times LikedD(u_a, u_b) \quad (3)$$

As the proposed metric is a composition of the two components, the liked and disliked dissimilarities are calculated using Eq. (4):

$$LikedD(u_a, u_b) = \sqrt{\sum_{i=1}^{n}((\frac{itemlikes_i}{maxLikes})^2 * (r_{u_a,i} - r_{u_b,i})^2)} \quad (4)$$

The liked dissimilarity function is characterized by its ability to add the global taste context to the dissimilarity computation between users. It measures how close the feedback of the users u_a and u_b is to the global positive taste, where it attributes more weight and importance to positive items and assesses how users u_a and u_b interact with these items. Similarly, the disliked dissimilarity function focuses more on less globally admired items [see Eq. (5)]:

$$DislikedD(u_a, u_b) = \sqrt{\sum_{i=1}^{n}((\frac{itemDislikes_i}{maxDisLikes})^2 * (r_{u_a,i} - r_{u_b,i})^2)} \quad (5)$$

where,

- $r_{u',i}$ are the ratings of user u' for item i.
- $maxLikes$ the maximum number of users who liked any single item.
- $itemLikes_i$ calculates the number of users who rated item i greater than or equal to the threshold RT.
- $maxDisLikes$ the maximum number of users who disliked any single item.
- $itemDislikes_i$ calculates the number of users rated item i less than or equal to the threshold RT.

To reflect the fact that humans differ in which side of taste they tend to express the most, we added the like and dislike weights [see Eq. (6) and Eq. (7)]:

$$weight_{like} = \frac{(\#likedItems_{u_a} + \#likedItems_{u_b})}{(\#ratedItems_{u_a} + \#ratedItemsu_b)} \quad (6)$$

$$weight_{dislike} = \frac{(\#dislikedItems_{u_a} + \#dislikedItems_{u_b})}{(\#ratedItems_{u_a} + \#ratedItemsu_b)} \quad (7)$$

where,

- $\#ratedItems_{u'}$: how many items were rated by the user u'.
- $\#likedItems_{u'}$: how many items were liked by the user u'.
- $\#dislikedItems_{u'}$: how many items were disliked by the user u'.

Therefore, the merged dissimilarity function, Eq. (3), is used to cluster users based on their preferences, Algorithm 1 defines the steps followed to cluster the users.

The final returned User Clusters, UC, group the grey sheep users candidate into clusters by similarity. In the next section, we explore how the final set of grey sheep users is identified.

Algorithm 1 Grey Kmedoids Clustering Algorithm

Require: Number of clusters K, Grey sheep candidate users $GC = \{gc_1, gc_2, \ldots, gc_n\}$
Ensure: UserCluster $UC = \{UC_1, UC_2, ..., UC_k\}$ and medoids $M = \{m_1, \ldots, m_K\}$

select K objects randomly from GC as initial medoids
$M \leftarrow \{m_1, m_2, \ldots, m_K\}$
repeat
  assignment:
  for each $gc \in GC$ **do**
    assign gc to the nearest medoid using

$$i \Leftarrow \arg \min_{j \in \{1,\ldots,K\}} d_{\text{merged}}(gc, m_j)$$

    add gc to cluster UC_i
  end for
  compute total cost:

$$S(M) = \sum_{i=1}^{K} \sum_{gc \in UC_i} d_{\text{merged}}(gc, m_i)$$

  swap:
  for each medoid $m \in M$ **do**
    for each non-medoid $o \in GC \setminus M$ **do**
      $M' \leftarrow (M \setminus \{m\}) \cup \{o\}$
      reassign all points using M'
      compute $S(M')$
      if $S(M') < S(M)$ **then**
        $M \leftarrow M'$
      end if
    end for
  end for
until medoids do not change
return M and clusters $UC_1, \ldots, UC_K$

C. Convergence Analysis of GTMedoids

As GTMedoids is a clustering method built based on Kmedoids, proving its convergence depends on proving that our algorithm can to reach a fixed point where neither assignments nor medoids change. In other words, the objective function should be monotonically non-increasing and cannot decrease indefinitely over a finite set [30].

Let $d_{merged}(\cdot, \cdot)$ denote the dissimilarity function used in the algorithm. The objective is the within-cluster merged-dissimilarity sum [see Eq. (8)]:

$$S'(\mathcal{UC}, \mathbf{m}) = \sum_{j=1}^{K} \sum_{i \in UC_j} d_{\text{merged}}(i, m_j), \tag{8}$$

where, $\mathcal{UC} = \{UC_1, \ldots, UC_K\}$ denotes the current user clusters, and $\mathbf{m} = (m_1, \ldots, m_K)$ denotes the set of cluster prototypes or medoids.

GTMedoids proceeds by alternating between two steps:

- **User assignment:** Given current medoid $\mathbf{m}^{(t)}$, each user i is assigned to the closest cluster UC [see Eq. (9)]:

$$uc^{(t+1)}(i) = \arg \min_{j \in \{1,\ldots,K\}} d_{\text{merged}}(i, m_j^{(t)}). \tag{9}$$

- **Medoid update:** Given the updated clusters $\mathcal{UC}^{(t+1)}$, each medoid is updated as Eq. (10):

$$m_j^{(t+1)} = \arg \min_{u \in UC_j^{(t+1)}} \sum_{i \in UC_j^{(t+1)}} d_{\text{merged}}(i, u). \tag{10}$$

The algorithm converges when neither the medoids nor the assignments change.

1) Proof of convergence:

a) Lemma a (User Assignment Step Monotonicity): Given fixed medoids $\mathbf{m}^{(t)}$. The assignment step does not increase the objective function S'.

Demonstration. Assigning user i to the cluster whose medoid minimizes the objective results in the smallest possible contribution for that user [see Eq. (11)]:

$$d_{\text{merged}}(i, m_{uc^{(t+1)}(i)}) \leq d_{\text{merged}}(i, m_{uc^{(t)}(i)}) \tag{11}$$

Hence, summing over all users, the total objective cannot increase [see Eq. (12)]:

$$S'(\mathcal{UC}^{(t+1)}, \mathbf{m}^{(t)}) \leq S'(\mathcal{UC}^{(t)}, \mathbf{m}^{(t)}). \tag{12}$$

b) Lemma b (Medoids Update Monotonicity): Given fixed cluster assignments $\mathcal{UC}^{(t+1)}$, the medoid update step does not increase S'.

Demonstration. For each cluster $UC_j^{(t+1)}$, the medoid $m_j^{(t+1)}$ is chosen as the minimizer of $\sum_{i \in UC_j^{(t+1)}} d_{\text{merged}}(i, u)$ over all $u \in UC_j^{(t+1)}$, as presented in Eq. (10).
By the characterization of the argmin, the chosen medoid cannot result in a larger value than any other one, including the previous one. Consequently, summing over all clusters yields [see Eq. (13)]:

$$S'(\mathcal{UC}^{(t+1)}, \mathbf{m}^{(t+1)}) \leq S'(\mathcal{UC}^{(t+1)}, \mathbf{m}^{(t)}). \tag{13}$$

c) Lemma c (Monotonic Decrease of the Objective): The objective sequence $\{S'^{(t)}\}$ generated by GTMedoids is monotonically decreasing.

Demonstration. Combining Lemmas a and b yields [see Eq. (14)]:

$$S'(\mathcal{UC}^{(t+1)}, \mathbf{m}^{(t+1)}) \leq S'(\mathcal{UC}^{(t+1)}, \mathbf{m}^{(t)}) \leq S'(\mathcal{UC}^{(t)}, \mathbf{m}^{(t)}). \tag{14}$$

d) Lemma d (Finite descent): GTMedoids terminates after a finite number of iterations.

Demonstration. Each medoid is selected from a finite set of user indices. Additionally, there are finitely possible partitions of n users into K clusters. Therefore, the number of possible combinations $(\mathcal{UC}, \mathbf{M})$ is finite.

Considering that the objective function is monotonically decreasing and cannot decrease indefinitely over a finite set, GTMedoids must eventually reach an optimized pair $(UC^{\star}, M^{\star})$ where neither the medoids nor the assignments change.

e) Theorem (Convergence of GTMedoids): GTMedoids converges in a finite number of updates to a local minimum of the objective function S'.

Demonstration. By Lemmas a–b, the objective function is monotonically decreasing.

By Lemma c, the sequence of solutions must reach an optimum; no additional changes are applied on the clusters, resulting in decreasing S' further.

D. Grey Sheep Users Identification

Clustering groups similar data points but struggles to clearly distinguish between dominant groups and outliers. In this regard, we have provided a new method that can be adopted for this matter in general, and not only for grey sheep user identification or recommender systems. Our technique is able to quantify the importance and weight of each cluster and detect the less impactful ones; in other words, we detect grey sheep users in our case.

1) Inter-Cluster Weights: We have successfully identified final clusters of users by analyzing their feedback and proximity based on multi-dimensional tastes and community preferences. The next step involves selecting only the grey sheep clusters, using the characteristics of GSUs as a baseline for this selection process:

- GSUs represent a minority; otherwise, mining their preferences and providing them with recommendations would have been easier.
- GSUs have distinguished preferences from users, meaning that a grey sheep user should be located far from all users.

We used the softmax function to calculate the weight of each cluster j, where the general equation is Eq. (15):

$$\tilde{\alpha}_j = \frac{exp(\alpha_j)}{\sum_{c \in UC} exp(\alpha_c)} \tag{15}$$

The softmax function heavily depends on α_j, to to obtain meaningful weights, we translated the GSUs characteristics to quantified metrics using the following Eq. (16):

$$\alpha_j = \begin{cases} \text{if } \#UC_j > 1: \\ (\frac{\#UC_j}{max(\#UC)}) \times (1 - \frac{(\sum_{gc \in UC_j}(MergedD(gc, c_j))}{((\#UC_j \times max(MergedD(UC_j, c_j)))}) \\ \text{otherwise:} \\ 0 \end{cases} \tag{16}$$

The GSUs representing the minority clause are represented by the first part of the α_j multiplication function, and the fact that a grey sheep user should be located far from all users is the second part. We attribute a zero impact to the cluster that we already know are grey sheep clusters as they have only one user. By applying the softmax to the α_j, we obtained inter-cluster powers $\tilde{\alpha}_j$.

2) Grey sheep clusters threshold: Having an ideal system clear of grey sheep users and each user is having enough similar users sharing with him the same preferences means that cluster powers should be equal, meaning all the clusters should have $\frac{1}{K}$ as power weight. However, as there is no such perfect system in reality, we have added a tolerance rate to not be too strict and handled normal users as grey ones just because they have some disagreements with the collaborative signals [see Eq. (17)].

$$\epsilon = \frac{Max(\tilde{\alpha}) - Min(\tilde{\alpha})}{K} \tag{17}$$

By including the clause ϵ to verify if the difference between the maximum and the minimum is equally distributed and provide a tolerance to make sure that only users with high derivation from the reset as identified as GSUs [see Eq. (18)].

$$T = \frac{1}{K} + \epsilon \tag{18}$$

This threshold can detect all the clusters that impact and influence more than the others, and then any cluster with a weight larger than this threshold should be categorized as a normal user cluster.

IV. Results and Discussion

In this section, we demonstrate the validity of our approach and evaluate its performance in identifying grey sheep users through experiments with two distinct datasets.

A. Evaluation Protocol and Metrics

Since there is no public benchmarking to tackle GSUs [13], [14], we assessed our methodology using the most commonly utilized datasets found in related research, MovieLens 100K (ML) [26] and FilmTrust (FM) [27], as shared in Table I.

To investigate the efficiency of our model in detecting GSUs, we employed the mean absolute error (MAE) as an evaluation metric. This metric quantifies the system's ability to reflect the user preferences. As a result, a GSU should have a high mean absolute error which demonstrates system failures in mining preferences. Moreover, we evaluatedthe MAE variation according to theuser type. We utilized the metrics introduced in [19], which denote the MAE increase

TABLE I. THE PROPRIETIES OF THE USED DATASETS

Dataset	#Users	#Items	#Ratings	Scale
MovieLens 100K (ML)	943	1682	1000	1-5
FilmTrust(FT)	1508	2071	35497	0.5-4

$Inc(MAE)$ for grey sheep users and MAE improvement $Imp(MAE)$ for normal users.

$$Imp(MAE) = 1 - \frac{MAE_{NU}}{MAE_{AllUsers}} \tag{19}$$

As GSUs preferences can also affect the overall quality of the system, theoretically excluding them from the recommendation should improve the suggestion. Eq. (19) quantifies the possible improvement for the remaining users in terms of MAE variation, where approaching 1 indicates the enhancement of the recommendation.

$$Inc(MAE) = 1 - \frac{MAE_{NU}}{MAE_{GSU}} \tag{20}$$

Eq. (20) was designed to evaluate the performance of the system in detecting grey sheep users, where values approaching 1 reflect the system's ability to identify these users. The MAE is obtained using Eq. (21), where UG is one of the user groups: all users, grey sheep users, or normal users.

$$MAE_{UG} = \frac{\sum_{(u,i)\in UG} |\hat{r}_{u,i} - r_{u,i}|}{\#UG} \tag{21}$$

To calculate the predicted ratings, we implemented User-Based Collaborative Filtering (UBCF) [28], as it highly depends on the nearest users to calculate the rating, which will help in assessing the performance of our model in detecting GSUs, where the ratings are predicted with:

$$\hat{r}_{u,i} = \tilde{r}_u + \frac{\sum_{(u')\in N_u} (r_{u',i} - \tilde{r}_{u'}) \times sim(u,u')}{\sum_{(u')\in N_u} sim(u,u')} \tag{22}$$

where,

- $\tilde{r}_u$ and $\tilde{r}_{u'}$: are the average ratings of the users u and u'.
- N_u: is the set of nearest users to the user u.
- $r_{u',i}$: is the rating r given by the user u' to the item i.

GTMedoids is based on an enhanced variation of the standard Kmeans clustering approach, in which the Euclidean distance is primarily used [see Eq. (23)]:

$$Euclidean(u_a, u_b) = \sqrt{\sum_{i=1}^{n} ((r_{u_a,i} - r_{u_b,i})^2)} \tag{23}$$

To demonstrate the need for the merged dissimilarity metric we compared the results obtained by the approach using Euclidean and our metrics. Where the user taste prediction is computed using the same technique in Eq. (22) with the Euclidean distance instead of the merged dissimilarity metric [see Eq. (24)]:

$$\hat{r}_{u,i} = \tilde{r}_u + \frac{\sum_{(u')\in N_u} (r_{u',i} - \tilde{r}_{u'}) \times Euclidean(u,u')}{\sum_{(u')\in N_u} Euclidean(u,u')} \tag{24}$$

In addition to all the metrics cited above, we introduce the average rating variance metric, where the rating variance of an item i is calculated as Eq. (25):

$$Variance_i = avg(r_{GSU,i}) - avg(r_{NU,i}) \tag{25}$$

where,

- $avg(r_{GSU,i})$ is the average rating of all the grey sheep users GSU who interact with item i.
- $avg(r_{NU,i})$ is the average rating of all normal users NU who interact with item i.

The $Variance_i$ metric is able to quantify the deviation of the grey sheep users' preferences from the normal ones, where values approaching 0 indicate a similar taste, while values approaching the maximum rating option represent a high deviation in the taste.

B. Experiments Setup and Configuration

To differentiate between positive and negative feedback, an average rating serves as a threshold favoring positive reviews, as ratings often indicate user interest, according to existing literature. The rating thresholds are set at 2.5 for FilmTrust and 3 for MovieLens. Furthermore, since an improved version of the Kmeans algorithm is part of this approach, it is crucial to determine the number of clusters (K) before executing the clustering algorithm. The elbow method was employed to identify the optimal number of clusters.

Based on the experimental results, the optimal K for $MovieLens100K$ and $FilmTrust$ is 3, as illustrated in Fig. 3. Therefore, for each dataset, we clustered the users into three clusters.

C. Results Analysis

Our model identified 273 of 943 users and 292 of 1508 users as grey sheep candidates on Movielens and FilmTrust, respectively, as shown in Table II, where 121 and 129 of them were detected as grey sheep users grouped in two grey clusters for both datasets.

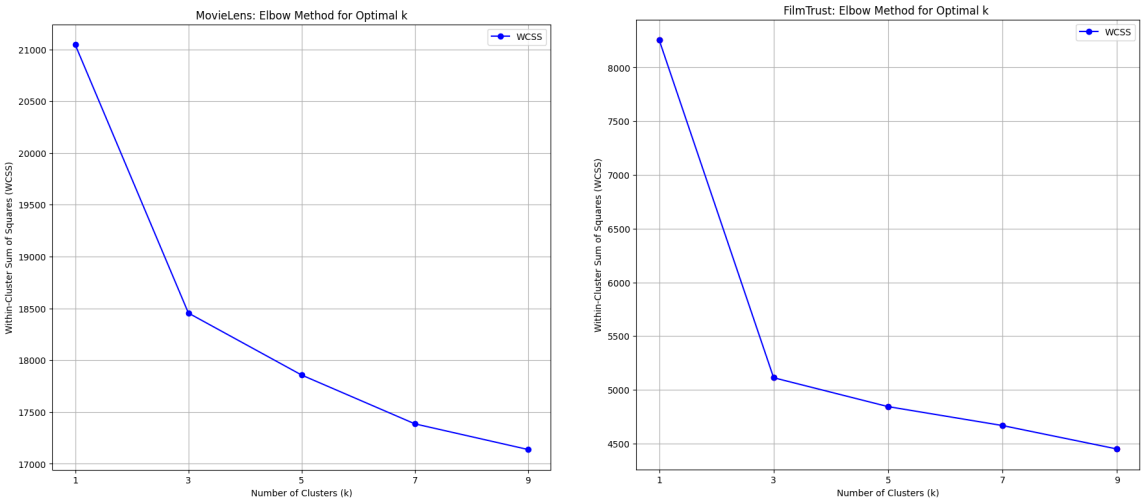


Fig. 3. Elbow method for identifying optimal K.

TABLE II. GREY SHEEP USERS DETECTION PER DATASET

Dataset	#users	#clusters	#grey Clusters	#GS candidate	#GSUs
ML100k	943	3	2	273	121
FM	1508	3	2	292	129

In assessing the validity of results, the impact of closest users on individual recommendations was measured. Grey sheep users exhibit high MAE values due to their closest users being more distantly located compared to the neighbors of normal users. This results in grey sheep consistently receiving inaccurate suggestions, as illustrated in Fig. 4, which compares MAE across user types.

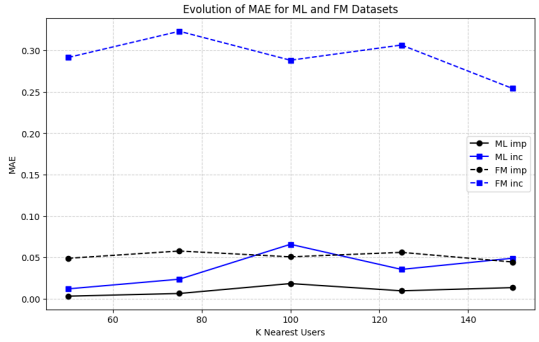


Fig. 4. MAE variations by user type.

The best results were achieved for K=75 and K=100 for FM and ML, respectively, where the mean absolute error almost doubled for the FM grey sheep users and increased by 6.616% for ML, as shown in Table III.

We propose GTMedoids, a clustering approach that categorizes users based on their greyness degree, contrasting with traditional methods that enhance centroid selection. Our model's efficiency is demonstrated through a comparison with existing literature [19], and the results are presented in Table IV. Note that FilmTrust has shown better outcomes, which can be justified by the RT precision as the rating scale is 0.5 and also the more valuable number of user-item interactions provided in this dataset.

Our approach has surpassed existing solutions in detecting grey sheep users while also enhancing recommendations for normal users, despite its primary focus on grey sheep selection within the community. This effectively supports the rationale behind our concept. Additional analysis was conducted by comparing average ratings of items rated by both user types, results of which are illustrated in Fig. 5. In the figure, red data points represent the average ratings given by grey sheep users, while blue data points reflect the overall average ratings from normal users for the same items. The figure clearly highlights the differences in opinion between the user types, evidenced by the distinct data point distributions for each item.

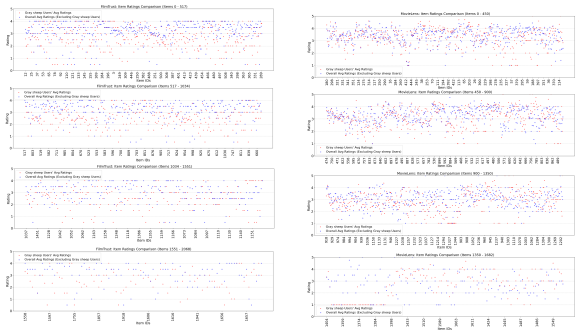


Fig. 5. Grey sheep and normal users average rating comparison.

To assist with the complexity of numerous data points, we calculated the average rating variance between grey and normal sheep users using Eq. 25.

The fluctuation presented in Fig. 6 illustrates the grey sheep user preferences deviation from the global collaborative taste and proves that these users rarely agree with the rest of the users when the difference is zero.

[1]Note: The results shared in this table were taken from the original work [19].

TABLE III. GTMEDOIDS' PERFORMANCE IN TERMS OF MAE VARIATION BY USER TYPE

	All Users	Normal Users	GSUs	$MAE_{NU}\%$	$MAE_{GS}\%$
MAE (FM) k=75	0.6776	0.6385	0.9434	-5.77%	$+32.31\%$
MAE (ML) k=100	0.8488	0.8332	0.8919	-1.83%	$+5.07\%$

TABLE IV. THE PERFORMANCE OF GTMEDOIDS IN TERMS OF GREY SHEEP USERS DETECTION

Methods	$Inc(MAE)$ ML	$Imp(MAE)$ ML	$Imp(MAE)$ FM	$Inc(MAE)$ FM
KMeans Standard[1]	0.027044	0.008989	0.014441	0.090519
KMeans++[1]	0.030882	0.007733	0.013397	0.089948
KMeans Pluspower[1]	0.037215	0.009111	0.018492	0.088580
Power Weight + Distance[1]	0.041141	0.011093	0.021348	0.093265
Power Weight Item + Distance[1]	0.029595	0.010316	0.020630	0.098378
GTMedoids	**0.065814**	**0.018378**	**0.057703**	**0.323192**

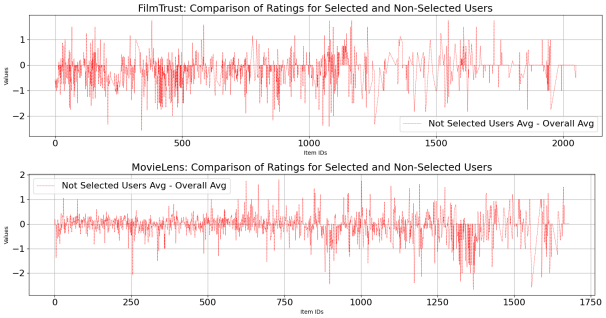


Fig. 6. The variation of ratings between grey sheep and normal users.

As we proposed a new dissimilarity function, we studied the impact of it instead of using Euclidean distance. To run the model using the traditional approach, we must determine the optimal number of clusters. As illustrated in Fig. 7, K was 3 for both ML and FM datasets.

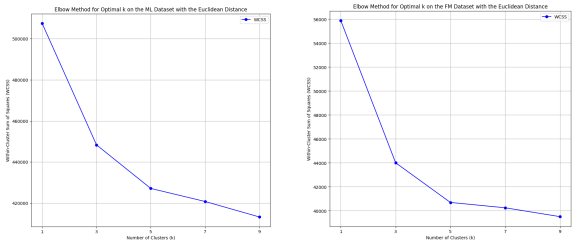


Fig. 7. Elbow method for identifying optimal K with the Euclidean distance.

By implementing the approach using the Euclidean distance, we obtained 155 and 142 users selected as GSUs in FM and ML, respectively. Fig. 8 shows the importance considering the feedback context while detecting GSUs in the system. Where the model was more accurate with our newly presented function such that not only fewer users were handled as GSUs but also the MEA increased. The selected users with the merged dissimilarity metric were also tested with the standard Euclidean function to predict their preferences. However, the prediction was less accurate for these users compared to the users detected and tested using only the Euclidean function. Therefore, the merged dissimilarity metric is more precise in detecting users who represent a minority in terms of taste and preference uniqueness.

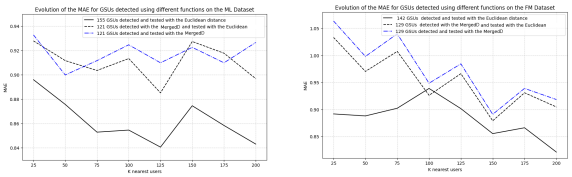


Fig. 8. The merged dissimilarity metric efficiency in detecting GSUs in comparison with the Euclidean distance.

This section experimentally demonstrates the performance of our model and its efficiency in detecting grey sheep users in the system. GTMedoids was able to identify users with unique preferences that the system often failed to recommend to them accurately.

V. CONCLUSION

In this study, we introduce Grey Threshold Medoids (GTMedoids), a novel method for identifying grey sheep users. The initial phase of our approach involves filtering users based on their frequency of agreement with the community's preferences, specifically isolating those who infrequently agree and frequently disagree with prevailing tastes for subsequent clustering. This selection process acknowledges the complexity of individual tastes, characterized as bi-polar, meaning that interaction with a particular item does not guarantee user appreciation. Clustering is conducted via a novel personalized metric that evaluates the significance of both positive and negative feedback from users while assessing their similarities. However, the clusters generated by this method do not inherently indicate which should be classified as grey sheep clusters. To address this, we have devised a new selection mechanism applicable beyond the realm of recommender systems. We compute the power of each cluster based on grey sheep user characteristics and finally identify the less-influencing clusters as grey sheep clusters that group our grey sheep users. The experiments proved the validity and accuracy of the proposed approach

As GTMedoids is based on the rating matrix in the future, it can be integrated to detect grey sheep users in pre-existing

RSs in the literature so that they can receive special treatment while making recommendations. Another future direction is to improve GTMedoids itself and not to be limited to just GSU detection but also by adding the recommendation module to it to treat each group of users fairly, either by adding KG graphs as a supplementary source of knowledge to strengthen user and item embedding or by leveraging the attention mechanism in such a way that the model learning phase can focus effectively on what is important to these users.

REFERENCES

[1] W. Zhang, X. Zhang, H. Wang, D. Chen, A deep variational matrix factorization method for recommendation on large scale sparse dataset, Neurocomputing 334 (2019) 206–218. doi:10.1016/j.neucom.2019.01.028.

[2] N. Idrissi, A. Zellou, A systematic literature review of sparsity issues in recommender systems, Social Network Analysis and Mining 10 (12 2020). doi:10.1007/s13278-020-0626-2.

[3] T. Trinh, V.-H. Nguyen, N. Nguyen, D.-N. Nguyen, Product collaborative filtering based recommendation systems for large-scale e-commerce, International Journal of Information Management Data Insights 5 (1) (2025) 100322. doi:https://doi.org/10.1016/j.jjimei.2025.100322.

[4] A. Srivastava, P. K. Bala, B. Kumar, New perspectives on gray sheep behavior in e-commerce recommendations, Journal of Retailing and Consumer Services 53 (2020) 101764. doi:https://doi.org/10.1016/j.jretconser.2019.02.018.

[5] A. A. Kardan, M. Ebrahimi, A novel approach to hybrid recommendation systems based on association rules mining for content recommendation in asynchronous discussion groups, Information Sciences 219 (2013) 93–110. doi:10.1016/j.ins.2012.07.011.

[6] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, Item-based collaborative filtering recommendation algorithms, in: Proceedings of the 10th International Conference on World Wide Web, WWW '01, Association for Computing Machinery, New York, NY, USA, 2001, p. 285–295. doi:10.1145/371920.372071.

[7] G. Behera, N. Nain, Collaborative filtering with temporal features for movie recommendation system, Procedia Computer Science 218 (2023) 1366–1373, international Conference on Machine Learning and Data Engineering. doi:https://doi.org/10.1016/j.procs.2023.01.115.

[8] A. Tourani, H. A. Rahmani, M. Naghiaei, Y. Deldjoo, Capri: Context-aware point-of-interest recommendation framework, Software Impacts 19 (2024) 100606. doi:https://doi.org/10.1016/j.simpa.2023.100606.

[9] B. Boualaoui, A. Zellou, M. Berquedich, Knowledge graph-based recommender systems to mitigate data sparsity: A systematic literature review, International Journal of Interactive Mobile Technologies (iJIM) 19 (03) (2025) pp. 115–140. doi:10.3991/ijim.v19i03.49427.

[10] Y. Zheng, An empirical comparison of outlier detection methods for identifying grey-sheep users in recommender systems, in: L. Barolli (Ed.), Advanced Information Networking and Applications, Springer Nature Switzerland, Cham, 2024, pp. 62–73.

[11] B. Gras, A. Brun, A. Boyer, Identifying grey sheep users in collaborative filtering: A distribution-based technique, in: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 17–26. doi:10.1145/2930238.2930242.

[12] Y. Zheng, M. Agnani, M. Singh, Identification of grey sheep users by histogram intersection in recommender systems, in: G. Cong, W.-C. Peng, W. E. Zhang, C. Li, A. Sun (Eds.), Advanced Data Mining and Applications, Springer International Publishing, Cham, 2017, pp. 148–161.

[13] B. Boualaoui, A. Zellou, L. Ajallouda, Towards a new evaluation protocol for grey sheep user detection approaches, in: M. Kodad, O. Moussaoui (Eds.), The 3rd International Conference on Artificial Intelligence and Smart Applications (AISA'25), Volume 1: Artificial Intelligence, IoT, and Smart Applications, Springer Nature Switzerland, Cham, 2026, pp. 13–25. doi:10.1007/978-3-032-18716-1_2.

[14] R. Alabdulrahman, H. Viktor, Catering for unique tastes: Targeting grey-sheep users recommender systems through one-class machine learning, Expert Systems with Applications 166 (2021) 114061. doi:https://doi.org/10.1016/j.eswa.2020.114061.

[15] H.-P. Kriegel, M. Schubert, A. Zimek, Angle-based outlier detection in high-dimensional data, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 444–452. doi:10.1145/1401890.1401946.

[16] M. Ghazanfar, A. Prugel-Bennett, Fulfilling the needs of gray-sheep users in recommender systems, a clustering solution, in: 2011 International Conference on Information Systems and Computational Intelligence (18/01/11 - 20/01/11), 2011, event Dates: 18-20, January.

[17] M. A. Ghazanfar, A. Prügel-Bennett, Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems, Expert Systems with Applications 41 (7) (2014) 3261–3275. doi:https://doi.org/10.1016/j.eswa.2013.11.010.

[18] X. Amatriain, J. M. Pujol, N. Oliver, I Like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems, in: User Modeling, Adaptation, and Personalization, Springer, Berlin, Heidelberg, 2009, pp. 247–258. doi:10.1007/978 − 3 − 642 − 02247 − $0_2$4.

[19] R. Rashidi, K. Khamforoosh, A. Sheikhahmadi, An analytic approach to separate users by introducing new combinations of initial centers of clustering, Physica A: Statistical Mechanics and its Applications 551 (2020) 124185. doi:https://doi.org/10.1016/j.physa.2020.124185.

[20] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Computing Surveys 41 (3) (2009) Article 15, 58 pages. doi:10.1145/1541880.1541882.

[21] B. Kaur, S. Rani, Identification of gray sheep using different clustering algorithms, in: D. Goyal, A. K. Gupta, V. Piuri, M. Ganzha, M. Paprzycki (Eds.), Proceedings of the Second International Conference on Information Management and Machine Intelligence, Springer Singapore, Singapore, 2021, pp. 211–217.

[22] H. Jindal, S. Agarwal, N. Sardana, Powkmeans: A hybrid approach for gray sheep users detection and their recommendations, Int. J. Inf. Technol. Web Eng. 13 (2) (2018) 56–69. doi:10.4018/IJITWE.2018040106.

[23] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1) (1967) 21–27. doi:10.1109/TIT.1967.1053964.

[24] B. Gras, A. Brun, A. Boyer, Can matrix factorization improve the accuracy of recommendations provided to grey sheep users?, in: Proceedings of the 13th International Conference on Web Information Systems and Technologies - WEBIST, INSTICC, SciTePress, 2017, pp. 88–96. doi:10.5220/0006302700880096.

[25] A. Nowak-Brzezińska, I. Gaibei, How the outliers influence the quality of clustering?, Entropy 24 (2022) 917. doi:10.3390/e24070917.

[26] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (4) (Dec. 2015). doi:10.1145/2827872.

[27] G. Guo, J. Zhang, N. Yorke-Smith, A novel evidence-based bayesian similarity measure for recommender systems, ACM Trans. Web 10 (2) (May 2016). doi:10.1145/2856037.

[28] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, Grouplens: an open architecture for collaborative filtering of netnews, in: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, CSCW '94, Association for Computing Machinery, New York, NY, USA, 1994, p. 175–186. doi:10.1145/192844.192905.

[29] C. C. Aggarwal, P. S. Yu, Outlier detection for high dimensional data, SIGMOD Record 30 (2) (2001) 37–46. doi:10.1145/376284.375668.

[30] Y. Mo, R. Xing, H. Hou, A distributed higher-order k-medoids clustering algorithm for network partition, IEEE Transactions on Network Science and Engineering 11 (5) (2024) 4181–4193. doi:10.1109/TNSE.2024.3402383.

[31] M. Tiwari, M. J. Zhang, J. Mayclin, S. Thrun, C. Piech, I. Shomorony, BanditPAM: almost linear time k-medoids clustering via multi-armed bandits, in: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), 2020, Article 856, pp. 1–12. doi:10.5555/3495724.3496580.

A Proxy-Based DevSecOps Framework for Multi-Tier Web Applications

Abderrahim Rida, Abdelaziz Bakhil, and Ayoub Ait Lahcen
Engineering Sciences Laboratory-National School of Applied Sciences,
Ibn Tofail University, Kenitra, Morocco

***Abstract*—Multi-tier web applications face significant implementation challenges in DevSecOps, including tool integration complexity, automation gaps, and cultural resistance. This study presents a proxy-based DevSecOps framework grounded in a formal architectural decomposition that transforms classical $O(n \times m)$ pipeline integration complexity into additive $O(n_1 \times m_1) + O(n_2 \times m_2)$ components through the separation of static and dynamic execution contexts. The framework is instantiated for PHP-based multi-tier web applications deployed on AWS infrastructure using Terraform-managed Infrastructure as Code principles; all ecosystem coherence metrics, toolset analysis, and complexity computations are derived within this specific technology context, and generalization to other language ecosystems or cloud platforms constitutes a boundary condition discussed in Section V. Theoretical contributions include: 1) a formal proxy pipeline architecture with mathematical complexity analysis demonstrating that complexity reduction is guaranteed when tool ecosystem coherence exceeds 70%; 2) systematic tool integration using PHP-specific tooling, yielding a theoretical ecosystem coherence of 76.9%; and 3) theoretical validation addressing 18 out of 20 identified DevSecOps implementation challenges. Mathematical analysis theoretically predicts a 48.13% reduction in tool integration conflicts and a 61.9% toolset reduction relative to traditional monolithic pipelines through context separation. All quantitative figures presented above are theoretically derived predictions, not empirically measured outcomes. They are formalized as falsifiable hypotheses H1 through H5 in Section V, with empirical validation identified as the primary direction for future work.**

Keywords*—DevSecOps; proxy architecture; multi-tier applications; PHP; AWS; security testing; CI/CD pipeline optimization; ecosystem coherence

I. Introduction

Incorporating DevSecOps in multi-tier web applications also presents significant challenges to implementing, including cultural resistance, automation gaps, and tool integrations complexitiy [1], [2]. While several reference architectures have been published [3], [4], [5], none exist that are specific to multi-tier application scenarios, in which rich deployment workflows, heterogeneous layers of services, and interactions between databases present unique security testing needs that standard DevSecOps practices do not come close to fulfilling.

Today, most common DevSecOps frameworks have three chief weaknesses regarding multi-tier applications: 1) monolithic toolchains resulting in $\mathcal{O}(n \times m)$ integration overhead, where n is the pipeline stages and m is the security tools, 2) too little architectural separation of phases of SAST and DAST, and 3) missing purpose-built frameworks optimized for meeting multi-tier architecture needs, including database syncing, service composition, and security settings optimized for targeted envs.

The matrix [4] provides the different integration points and levels of automation associated with multiple DevSecOps tools throughout the software development lifecycle (SDLC). Tools such as open-source governance and Static Application Security Testing (SAST) show a high level of automation throughout the process. In the case of others, there is Dynamic Application Security Testing (DAST) and Interactive Application Security Testing (IAST) that are mostly automated in the Continuous Integration/Continuous Deployment (CI/CD) and Post-Deployment phases. Mobile application security testing and Runtime Application Self-Protection (RASP) at best automate a couple of capabilities. The hierarchical approach of adoption emphasizes how critical it is to begin with secure software development through automation and integrating testing at each level, hence making quality SDLC efficient and effective [6].

Report [3], [4], [5] is based on a list of tools intended to integrate security into the SDLC. Jenkins and Ansible provision infrastructure and CI/CD pipelines automatically, whereas Twistlock, Aqua Security, and Docker secure containerized applications and cloud-native platforms. To scan for vulnerabilities, the tools OWASP ZAP, Nessus, SonarQube, and Veracode perform static, dynamic, and interactive application security testing (SAST, DAST, IAST). Dependency management is handled by Nexus Lifecycle and Dependency-Check, which identify third-party dependency vulnerabilities. Istio also provides microservices' service mesh security, and Splunk, FireEye, etc., provide monitoring and threat detection capability. All these integrated together offer a robust DevSecOps pipeline in which security is integrated from code creation through to deployment and runtime [7].

Challenges identified in recent studies [2], [8], [9] highlight our motivation to create a specific CI/CD pipeline that could address some of these issues. Since the goal of DevSecOps is to promote quick and secure delivery, it is important to develop a guide that can help DevSecOps teams in their pipeline. This work further provides a reference implementation for the DevSecOps community. Right now, looking for existing pipelines will allow us to understand the details of each one and extract best practices from them. By taking this approach, we can design a useful pipeline by learning from existing work and finding solutions to many challenges.

- Design a proxy-based DevSecOps architecture that reduces tool integration complexity for multi-tier web applications.

- Develop a theoretical framework demonstrating mathematical complexity reduction through architectural separation.
- Provide a specialized solution addressing multi-tier web application deployment challenges.
- Establish a foundation for empirical validation of the proposed theoretical framework.

This research focuses specifically on multi-tier PHP web applications deployed on AWS infrastructure, utilizing Infrastructure as Code principles through Terraform implementation. The framework addresses static security testing separation while maintaining dynamic testing integration in production environments.

This study provides a proxy-centric DevSecOps framework to effectively overcome these limitations by employing architectural separation concepts tailor-made for multi-tier web applications. Our conceptual contributions include: 1) an accurate definition of the proxy pipeline architecture and an evaluation of its mathematical complexity, 2) an orderly procedure for tool integration based on ecosystem coherence metrics, and 3) an exhaustive conceptual verification showing the mitigation of challenges realized in DevSecOps implementation.

The manuscript is structured as follows: the next section contains some necessary background knowledge with the concepts and tools used in such pipelines. The third section presents our motivation and problem formalization. The fourth one examines the relevant worldwide DevSecOps pipelines. The fifth describes the specialized pipeline underlining the envisaged approach of this research. The last one provides the theory validation and Analysis of the proposed Multi-tier Web applications DevSecOps pipeline.

II. Related Work

A. Comprehensive DevSecOps Reference Architectures

The DevSecOps reference architecture collection [3] offers several pipeline deployments which illustrate a wide range of ways security is integrated into the SDLC. The architectures focus on a number of core principles such as continuous security integration, shift-left strategies whereby security tests are pushed into earlier stages of the SDLC, and automated security testing paradigms. The collection features deployments from basic CI/CD pipelines which include security tools integrated into the pipeline all the way through enterprise-class architectures which accommodate microservices environments.

In these reference architectures [3], various implementation patterns arise. The pipelines indicate trust and confidence development by security tool integration as the significant elements of CI/CD processes. The implementations compare DevOps delivery-based methods with DevSecOps security-included methods in the whole of the pipelines. The architectures indicate how the integration of security covers the whole SDLC activities from planning and code writing phases up through deployment, including the security as well as code analysis for potential vulnerabilities during the beginning of development cycles.

These reference architectures [3] also demonstrate workflow automation tactics employing tools including Jenkins, unit testing, code analysis, and package creation of artifacts. These pipelines deploy broad monitoring with feedback loops for the formation of continued improvement, indicating how security can be embedded across software supply chains where DevOps would traditionally focus on delivery teams whereas DevSecOps performs the security activities across full pipelines.

B. Sonatype DevSecOps Framework

Sonatype DevSecOps Reference Architecture [4], [5] outlines core pillars from ongoing education for every party at interest, secure supply chains for open-source components, and automated test methodologies using SAST, DAST, feedback mechanisms and monitoring. The framework includes detailed automation levels and integration points involving a variety of DevSecOps tools during the software development lifecycle.

The Sonatype architecture [4], [5] shows the potential of tools, such as Static Application Security Testing and open-source governance, to reach high levels of automation during the development lifecycle. The configuration shows clear integration of the security controls during the SDLC, whereby developers develop the applications, store them in the Nexus Repository, and use lifecycle management tools like Nexus Lifecycle for constructing and delivering. This approach ensures detection of vulnerabilities during the application build or release stages.

Under the reference implementations [4], [5], the architecture embodies holistic CI/CD pipelines for absorbing security requirements during the coding, building, testing, release, deployment, operation, and monitoring stages. Security check measures are integrated into each stage, which implies that security is not merely added but ingrained into each stage of software development. The framework highlights the importance of involving security teams throughout every stage, rather than having them only deliver analysis reports to the operations teams.

C. Formal DevSecOps Implementations

E-SPIN Corporation [10] represents the concept of Continuous Security within a DevSecOps framework, enabling a proactive approach to find vulnerabilities and integrating security into both development and operations tasks.

The DoD Enterprise DevSecOps [11] presents a technology stack that involves embedding and deploying security measures and practices into the SDLC with various tools. Each component has the goal of securing and automating the software delivery pipeline with appropriate DoD compliance requirements. The DoD Enterprise DevSecOps Platform Architecture details a secured microservices architecture showing how Istio creates a Service Mesh architecture for managing microservices, bringing traffic management associated with security features to address complex microservices environments.

The Application Security Orchestration and Correlation framework demonstrates how diverse security tools come together to execute assessments and identify vulnerabilities, ensuring effective security governance throughout the SDLC [12].

D. Industry DevSecOps Pipeline Implementations

Industry-based DevSecOps pipeline executions [13], [14] target software supply chain security and how critical effective dependency management and vulnerability scanning during the SDLC ensure no vulnerable applications get shipped. CI/CD pipelines with a security focus include all the security-related activities so pipelines do not overlook critical security concerns at each stage of software release, leading to more reliable and secure releases.

Current DevSecOps practices [15] are formulated as frameworks aimed at safeguarding the software supply chain, illustrating how instruments such as Dependency-Check and Dependency-Track can detect and monitor vulnerabilities within software dependencies. This approach helps organizations keep their software secure by carefully managing all its dependencies.

Further industry deployments [16] provide DevSecOps automation orchestration pipelines for software development lifecycle automation with security emphasis. These strategies exhibit how Information Security teams work together with IT Operations teams for infrastructure automation code review and contribution purposes, providing security integration into the components of the infrastructure.

E. Evolution of Security-Integrated Development

Software development has evolved from traditional, isolated methods to more integrated approaches that seamlessly incorporate security throughout the process. This progress reflects the interweaving of security concepts during development phases from planning and design all the way through ongoing tests and monitoring in production settings.

Today's DevSecOps platforms come with collections of tools which automate and simplify the processes engaged with software development. Such deployments enable development teams to improve productivity, communicate better, and deploy software of high quality with increased security mitigations. Distributed build and test operations are also facilitated by these architectures, thus meaning security has been integrated at multiple points, from testings in a range of environments to further security analyses of the CI/CD pipelines themselves.

Today's secured CI/CD deployments indicate the value of security throughout the SDLC in the form of secure delivery, which serves as the prime objective of DevSecOps architecture integration. These strategies indicate how security tools may be packaged within central platforms providing consistency as well as automation for security services necessitating coordinated responses.

F. Research Gap Analysis

While current DevSecOps reference architectures [3], [4], [5] offer detailed guidelines for integrating security, they don't fully address the challenges of applying these practices to multi-tier web applications.. They tend to treat deployment as monolithic activities without the specific needs for database sync, service orchestration, and tier-specific security requirements which characterize multi-tier applications.

The compared architectures support a set of methodologies for integrating tools but lack well-structured frameworks for managing the complexity involved when coordinating a set of multiple security tools operating at different application layers. The current deployments also don't provide an architectural separation between the static and dynamic security testing phases, leading to resource conflicts and less than optimal running of the pipeline.

This study tackles these deficiencies by introducing a proxy-based architecture that is tailored for multi-tier web applications. It offers a theoretical framework aimed at diminishing complexity through architectural segregation, all while ensuring extensive security coverage throughout the various application tiers.

III. Problem Formalization

A. DevSecOps Implementation Challenges

Analysis of existing DevSecOps resources [3], [4], [5] reveals that , the context of DevSecOps pipelines suggests that there are significant barriers to introducing security into SDLC. Challenges [2], [8], [9] include tool integration, automation gaps, cultural resistance, compliance, and limited resources, among others. Addressing these challenges entails technical solutions, culture change, and continuous improvement.

Various tools like Jenkins, Twistlock, Aqua, OWASP ZAP, and others are mentioned and need to be combined, this leads to Integration Complexity , because Integrating multiple security tools into the CI/CD pipeline can be complex and may lead to tool conflicts or inefficiencies.

Dynamic Application Security Testing and Run-time Application Self Protection are not fully automated across all stages of the pipeline. To achieve full automation of security testing and compliance checks is difficult.

The activities of identifying, prioritizing, and remedying vulnerabilities within a reasonable timeframe can be challenging, particularly with DevSecOps environments. Metrics such as "Time to patch vulnerabilities" reflect the importance of quick detection and remediation of vulnerabilities that are frequently challenging.

SecDevOps pipelines implement a number of tools at various stages that could swamp operations, security and developers teams. Integrations of those numerous security tools can bring about a state of tool overload as teams become challenged with managing and maintaining the tools efficiently.

Cultural aspect is highlighted through collaboration and mutual responsibility of DevSecOps Teams that are hard to realize across most organizations. DevSecOps needs a cultural approach where the teams of development, security, and operations work intimately. In the same manner, the requirement of "Security Champions" as an over-reliance on specialty skills. There just are no trained individuals with dual security and DevOps skill sets available to implement and run DevSecOps pipelines.

Compliance validation and policy enforcement can cripple the pipeline process if not properly established. Being compliant with regulatory requirements and being agile at the same time in the pipeline is a very big challenge.

Static Application Security Test are prone to false positives that need to be manually validated. Security tools also generate

false positives, which can lead to alert fatigue and lower the effectiveness of the pipeline.

DevSecOps pipelines reference the difficulty of scaling security in big distributed systems especially in micro-services architectures.

Secured pipelines will be likely centered on centralized logging and telemetry, signaling that visibility is concern number one. Visibility across the security stance of applications and infrastructure end-to-end is challenging to obtain, especially in cloud-native environments.

Some of the biggest challenges lie in managing and securing third-party dependencies (e.g., open-source libraries), as security vulnerabilities in dependencies can lead to risks.

Real-time monitoring and feedback are highlighted, and this requires good processes and tools. Establishing a continuous feedback loop to improve security practices over a period of time is tricky, especially in changing environments. On the same line, The DoD architecture has Zero Trust as one of the Istio benefits, but implementing it is hard. Applying a Zero Trust model to micro-services architecture is hard and expensive.

Legacy systems are not necessarily part of modern DevSecOps pipelines: It is difficult to incorporate DevSecOps practices into legacy systems and applications due to old architectures and the lack of automation.

Commercial tools with high licensing fees and maintenance are used in many secured pipelines, It is financially costly to put in place a DevSecOps pipeline, having high tool investments, training, and infrastructure costs.

Addressing security incidents in a timely manner in a DevSecOps environment requires a strong process and tools, and that may be difficult to effect.

Metrics like "Deployment frequency" and "Change lead time" identify the tension between velocity and security.Where achieving speed of delivery is an imperative, and one also needs to ensure thorough security testing, this is a constant challenge in DevSecOps.

Some tools are configured with no integration plan in sight: having multiple disparate tools creates fragmentation, where it is difficult to get an end-to-end integrated view of security across the pipeline.

Threat modeling tends to be an afterthought in high-velocity environments, including it in the DevSecOps pipeline is not straightforward because it requires skill and time.

The importance of constant improvement is always required, but this is an ongoing effort and cost. Maintaining pace with changing security threats and adjusting the pipeline in response is an ongoing challenge. Table I highlights the most important DevSecOps challenges [2], [8], [9].

B. Problem Formalization for Multi-Tier Applications

Multi-tier web application DevSecOps has mathematically expressible complexity challenges. Assume $P = \{p_1, p_2, \ldots, p_n\}$ to be a DevSecOps pipeline with n stages, and $T = \{t_1, t_2, \ldots, t_m\}$ to be m tools. The classical methods establish dependency relations $D = \{(p_i, t_j) \mid$ it executes as tool t_j in phase $p_i\}$, leading to pipeline complexity $O(n \times m)$.

Multi-tier applications multiply this complexity through:

- Requirements for synchronizing databases across staging, production, and development environment.
- Service level dependencies requiring sequenced deployment synchronizations (service orchestration).
- Environment-specific configuration management and security policy enforcement.
- Cross-tier communication security verification criteria.

In this particular instance, service orchestration means controlled management of several layers of an application whereby:

- Database tier must be operational before you can install application tier.
- Application services have to reach healthy state prior to web tier configuration.
- Security policies should be implemented and verified in a sequential manner at each level.
- Failure at one level requires coordinated rollback in precedent levels.

Proxy architecture overcomes these challenges through separation of execution contexts:

$$\text{Static Context } C_s = \{\text{checkout, build, compile, unit_test, code_analyze, style_test, security_scan}\},$$

$$\text{Dynamic Context } C_d = \{\text{deploy, monitor, runtime_protection}\}.$$

The separation reduces complexity from multiplicative

$$O(nm)$$

to additive

$$O(n_1 m_1) + O(n_2 m_2),$$

where, $n_1 + n_2 \leq n$ and $m_1 + m_2 \leq m$.

This achieves theoretical reduction whenever tool ecosystem coherence exceeds 70%.

IV. Our Proposal Proxy DevSecOps Architecture

A. Architectural Foundation and Benefits

These documented DevSecOps architectures have significant benefits from baking security directly into the SDLC. This "shift-left" approach improves overall security posture through automated security testing, continuous monitoring, and earlier detection of vulnerabilities for a more robust product with fewer security breaches. Again, a focus on automation, continuous integration and delivery, and collaboration between

TABLE I. DEVSECOPS CHALLENGES

DevSecOps Challenge	Brief Description
Tool Integration	Difficulty combining multiple security tools without conflicts.
Automation Gaps	Security tools not fully automated across the pipeline.
Patch Management	Hard to identify, prioritize, and fix vulnerabilities quickly.
Tool Overload	Too many tools overwhelm Dev, Sec, and Ops teams.
Cultural Resistance	Lack of collaboration and shared responsibility between teams.
Skills Gap / Security Champions	Shortage of professionals skilled in both security and DevOps.
Compliance Constraints	Balancing agility with regulatory compliance is difficult.
False Positives	Manual validation of false alerts leads to inefficiency.
Scalability	Difficult to scale security across distributed/microservices systems.
Limited Visibility	Hard to get end-to-end insight across cloud-native infrastructure and apps.
Third-Party Dependencies	Risks from vulnerabilities in open-source and external components.
Continuous Feedback Loop	Hard to build effective, timely feedback systems in fast-paced environments.
Zero Trust Implementation	Complex and costly to apply Zero Trust models to microservices.
Legacy Systems Integration	Difficult to apply DevSecOps to outdated systems lacking automation.
High Costs	Licensing, infrastructure, and training for DevSecOps tools are expensive.
Incident Response & Root Cause Analysis	Demands strong coordination and effort to resolve security incidents.
Velocity vs. Security	Balancing fast delivery with deep security testing is a continuous tension.
Tool Fragmentation	Disparate tools without integration reduce pipeline visibility.
Threat Modeling Neglect	Often skipped due to time and expertise required.
Continuous Improvement Burden	Constant updates needed to keep pace with evolving threats.

the development, security, and operations teams yields huge gains in efficiency and speed. Smaller feedback cycles, less rework, and better distribution of resources mean shorter time to market and costs for the firm, while affording the ability for the organization to adapt quicker to change. However, in practice, the adoption of these DevSecOps architectures brings considerable challenges in implementation: huge cultural change resisted within organizations, extensive training in new tools and processes, security automation tooling and technology may be expensive and complex to implement, requiring careful selection, seamless integration, and ongoing maintenance.

W3Techs conducted research showing that PHP is the top server-side programming language, with a usage rate of 74.0%. Nginx is the most commonly used web server, making up 33.7% of the market. Unix is the leading operating system, with a notable 89.3% share. Meanwhile, Amazon is the most popular data center provider, holding 9.5% of the market. This information gives us a basic understanding of the environment for our proposed DevSecOps pipeline. The objective of this work is to design a secure, automated pipeline.

B. Mathematical Framework and Formal Definitions

Definition 1 (Proxy Pipeline): A proxy pipeline P_p is defined as a 6-tuple (S, T, D, R, V, M) where:

- **Static analysis stages:** $S = \{$checkout, build, compile, unit_test, code_analyze, style_test, security_scan, web_performance_test$\}$
- **PHP Tool set:** $T \subseteq T_{\text{PHP}}$, where `php-composer, php-compiler, phpunit, phpca, phploc, phpsa, testability, pdepend, phpcf, phpdd, phanalist, phpmd, phpcs, phpcbf, phpcpd, phpdcd, parallel-lint, local-php-security-checker, php-scanner, php_upgrade`
- **Proxy pipeline deployment operations:** $D = \{$deploy, compile, database_refresh, database_insert, static_monitoring$\}$
- **Isolated compute resources:** $R = \{$EC2_dev, EC2_production, EC2_jenkins$\}$
- **Validation function:** $V : S \times T \rightarrow \{$pass, fail, warning$\}$
- **Static Monitoring results:** $M : S \times D \times R \rightarrow \{$cpu_usage, memory_usage, I/O_usage$\}$

Definition 2 (Tool Ecosystem Coherence): For a tool set T_{PHP}, ecosystem coherence is defined as:

$$EC_{\text{PHP}} = \frac{|T_{\text{PHP}} \cap T_{\text{required}}|}{|T_{\text{required}}|}$$

For our implementation: $EC_{\text{PHP}} = 20/26 = 0.769$ (76.9% coherence)

Theorem 1 (Complexity Reduction): The proxy architecture achieves complexity reduction $C_{\text{proxy}} < C_{\text{traditional}}$ when ecosystem coherence $EC > 0.7$.

In order to empirically prove the hypothesis of reducing complexity, we compared three pipeline instances derived from our implemented proxy architecture. The proxy pipeline (static context) consists of $n_1 = 31$ stages running $m_1 = 26$ SAST tools with a complexity of 806 operations. The production pipeline (dynamic context) has $n_2 = 9$ stages running $m_2 = 18$ DAST tools such as OWASP ZAP, Nikto, Burp Suite, w3af, Arachni, SQLMap, XSSHunter, LoadRunner, New Relic monitoring, as well as runtime protection tools (Invicti, OpenRASP), leading to 162 operations. A traditional monolithic pipeline combining both SAST and DAST approaches on the other hand will need $n = 37$ stages running $m = 42$ tools with a total of $1,554$ operations. The context separation leads to achieving 806 operations in the proxy context with a reduction in complexity by a percentage of 48.13% compared to the traditional approach $((1,554 - 806)/1,554)$. Our empirical comparison thus proves our theoretical proposition that breaking down DevSecOps complexity from $O(n \times m)$ down to $O(n_1 \times m_1) + O(n_2 \times m_2)$ by context separation leads to significant improvements in operational efficiency without

sacrificing complete coverage across phases of static as well as dynamic testing.

C. Proxy Pipeline Architecture Design

Our contribution will be iteratively improved by reaching maturity in successive versions. One of the peculiarities of our approach is the proposed pipeline, which may act as a proxy for the DevSecOps Architecture.

To explain the seminal idea of this proxy pipeline, it is necessary to describe its structure. The detailed explanation of all the visible steps is as follows:

- **Start:** The beginning of the pipeline execution.
- **Checkout Code:** This stage retrieves the latest code changes from the version control system (likely Git).
- **Building on dev:** This stage compiles and builds the code in a development environment.
- **Compiling on dev:** Another stage specific to the development environment.
- **Unit Testing:** The code is run through its unit tests to ensure the logic works as expected.
- **Analyzing Test:** Several automated tests run to analyze the code.
- **Styling Test:** Several automated tests run to analyze styling aspects of the code.
- **PHP Security:** Automated security analysis tools are executed for PHP.
- **Web Performance Test:** Testing focused on measuring the performance of the web application.
- **Deploying:** The code is deployed to a staging or preproduction environment.
- **Building on prod:** The code is built in a production environment.
- **Compiling on prod:** The code is compiled in the production environment.
- **Database Refresh:** The database is updated with the new schema or data.
- **Database Data Inserting:** The database is inserted with data.
- **End:** The completion of the pipeline.

D. Infrastructure Implementation Strategy

To provide a clear picture of the chosen infrastructure, let's discuss how it was developed. We selected Amazon Web Services (AWS) as a leading cloud provider based on W3Techs. Given its many benefits, we planned to use Infrastructure as Code (IaC) because it allows us to automate managing the infrastructure. This approach improves resource allocation [17], saves time, speeds up deployment, reduces human and manual errors, offers flexibility for backups and duplicates, and ensures reliability for secure CI/CD.

The study concerns the use of a multi-tiered PHP web application that is built on a commercial CMS. The application was deployed in a production environment on AWS with all services required for production environment management. The framework is designed to ensure that each commit is validated prior to affecting the production environment. Our intention is to provision three EC2 instances hosting both Development and Production environments, and a Jenkins instance, using the DevSecOPs pipeline model. By implementing the DevSecOps model we can establish process flows that prevent release regressions before we publish. The pipeline we have established is initiated with a Terraform file that produces the necessary EC2 instances, implementing the principles of Infrastructure as Code.

E. Pipeline Characteristics and Tool Integration

The given observations revealed the following characteristics:

- Automated Workflow: This is a fully automated pipeline.
- PHP-Based Application: The presence of "PHP" throughout many stages suggests that this pipeline is for a PHP-based application.
- Security Focus: The inclusion of security checks ("PHP Security") emphasizes the importance of security in this development process.
- Comprehensive Testing: The different stages from unit, analysis, styling, security, and performance testing show a focus on quality code.
- Database Component: The pipeline includes database refresh and database data inserting, suggesting the application interacts with a database.

It is noticeable that the majority of the tools used in our proposed pipeline are associated with PHP, therefore facilitating its implementation.

F. Proxy Pipeline Operational Framework

It's a full proxy CI/CD pipeline for a PHP-based application, with automation from check-out to database updates and deployment. It fetches code, builds, and compiles in the development environment, followed by rigorous unit, analysis, and styling tests using tools particular to PHP. The workflow also prioritizes security with dedicated security checks and web performance before moving to deployments in production environments that include production build, compilation, database refresh, and database inserts. This fully automated pipeline really shows a commitment to quality, security, and performance, emphasizing efficient and continuous delivery of the application.

This proposal pipeline aims to be simpler to implement with a guarantee of fast and secured delivery. Stages correlated and specifically detailed for PHP multi-tier web applications help organizations quickly adjust the steps according to their needs. This proxy pipeline is an exact balance between the efficiency of DevSecOps tools and their complexity in terms of integration.

The following subsection details the operational sequence of the proposed pipeline. We assume that our multi-tier web application was developed for research purposes, and it's running correctly on AWS with the services it needs. Additionally, for any new commit, we launch three EC2 instances by running a Terraform file; once the instances are started, the proxy DevSecOps process is subsequently initiated. Our pipeline is based on many Blocks such as:

1) First block: checkout starts the pipeline, followed by building and compilation in order to check if the code is functional, optimized, and with errors minimized, ready for deployment;

2) Second block: The unit, Analyzing, Styling tests represent a pre-phase of the security implementation. They are based on lightweight PHP tools, and their goal is to guarantee the quality, reliability, maintainability, and correctness of the PHP codebase. Each test type plays a particular role in the software development life cycle, and together they help to deliver a robust and well-structured application;

3) Third block: This is the security step from PHP, where it tries to find vulnerabilities, stopping the pipeline in case some security issues will be revealed;

4) Fourth block: In this context, post-phase security verification has the role of verifying whether the web server performs well enough in case everything is handled appropriately upon request;

5) Fifth block: if the commit has passed correctly the previous blocks, the deployment step is occurred and the building and compiling operations are done;

6) Sixth block: in order to verify that, in case some configurations or data have been added or modified in the development environment, it's necessary to transfer them onto the production environment.

G. Proxy Architecture Innovation and Separation Strategy

The three jobs can be disabled and closed once the commit has been successfully aligned with the pipeline. A simplified pipeline can be created after explaining this new idea of a proxy pipeline, which is supposed to run in the background. It will include the following stages: code checkout, building and compilation in the actual production environment, deployment, refreshing and inserting in the database, and monitoring.

The advantage of this innovative concept is that it will spread the tasks within the pipeline more efficiently by moving part of the operations to the background, keeping only the main stages. With this architecture, we guarantee that the production environment works in a safe mode. On the other hand, within the Proxy pipeline, we are implementing only Static Application Security Testing, while in the real production environments, we keep Dynamic Application Security Testing.

The models we are talking about in this work assume that tools are used in a straightforward way. Each part of the process is completed before the next part starts and the tools always behave the way.. In real life things are not that simple. The tools may be used at the time and they may be started and stopped at different times. This can cause delays between the parts of the process that we cannot predict.

The way we break down the complexity of the process into parts like $O(n_1 \times m_1).O(n_2 \times m_2)$ is still valid even if the tools are not used in a way. This is because the different parts of the process are kept separate so they do not affect each other. However the time it takes to complete the process that we predict with Theorem 5 is the best we can do when the tools are used in a way. It is not a guarantee of how it will take when the tools are used in a more complicated way.

We think it would be an idea to make our models more realistic by taking into account that the time it takes to complete each part of the process can vary. This would make our models more useful, for life situations and it is something we would like to work on in the future.

H. Challenge Resolution Analysis

This outlines our envisioned pipeline to accommodate static tests silently in the background to ensure the pipeline functions appropriately before administering any dynamic tests with monitoring in the true production environment. This is why we refer to our pipeline as a Proxy DevSecOps pipeline. In addition, the trade-offs illustrated suggest we provide added value as we alleviate both challenges as highlighted in the Table II.

V. Theory Validation and Analysis

A. Complexity Analysis and Foundations of Mathematics

1) Classical DevSecOps pipeline complexity model: Conventional DevSecOps deployments for multi-tier web applications display multiplicative complexity featuring tool-stage interdependencies. According to extensive modeling of available DevSecOps architectures [3], [4], [5], an average complete pipeline features n stages as well as m security development tools resulting in $O(n \times m)$ integration complexity. Multiplicative nature of the model presents the following operational challenges: 1) permanent resource allocation within the whole pipeline lifecycle irrespective of the utilization behavior, 2) cascading propagation of the failures between interdependent tool chains where individual tool breakdown may invalidate the whole pipeline realization, and 3) linear growth of complexity when adding individual tools or stages resulting in prohibitive costs for complete security coverage.

2) Proxy architecture decomposition framework: Our proxy-based architecture basically transforms this complexity structure through architecture separation into two independent execution contexts. Their underlying theory is based on decomposing the standard $O(n \times m)$ complexity into additive components $O(n_1 \times m_1) + O(n_2 \times m_2)$ where contexts execute independently without cross-dependencies.

Theorem 1 (Complexity Decomposition): Given a monolithic DevSecOps pipeline $P_{\text{traditional}}$ of n stages and m tools, and also given the proxy architecture of static context (n_1 stages, m_1 tools) and production architecture of dynamic context (n_2 stages, m_2 tools), then:

$$C(P_{\text{traditional}}) = O(n \times m),$$

where, $n \geq n_1 + n_2$ and $m \geq m_1 + m_2$

TABLE II. RESOLUTION OF DEVSECOPS CHALLENGES USING THE PROXY PIPELINE ARCHITECTURE

DevSecOps Challenge	Theoretical Resolution Mechanism	Supporting Evidence	Confidence Level
Tool Integration	PHP ecosystem coherence ($> 70\%$) achieving reduced dependency conflicts	Mathematical analysis proves a 48.13% reduction in conflict via homogeneous instrument choice	High
Automation Gaps	Complete CLI automation for all of the 20 PHP tools in proxy pipeline	All selected tools support automated execution without manual intervention	High
Patch Management	Vulnerability detection using PHP security tools	Automated proxy stage vulnerability detection prevents deployment of vulnerable components	High
Tool Overload	Tool set reduction of 42 (traditional) to 26 (proxy) as a 61.9% reduction	Tool number analysis indicates streamlined management overhead	High
Cultural Resistance	Shared PHP tooling lowers learning curve and adoption costs	Capitalizes on current development team PHP skills and competency	Medium
Skills Gap / Security Champions	DevSecOps-specific tools demand a current skill set instead of specialized security expertise	Decreases reliance on few DevSecOps specialists	Medium
Compliance Constraints	Automated checking for compliance using integrated PHP security tools	Compliance validation for regulation provided by PHP tools	Medium
False Positives	Tool-specific optimization mitigates cross-tool conflict and false alarms	Ecosystem consistency for PHP decreases interactions between tools	High
Scalability	Autonomous context scaling via architectural segregation	Provable mathematical evidence shows additive, not multiplicative complexity	High
Limited Visibility	Unified logging and monitoring throughout both proxy and production environments	Central monitoring system provides end-to-end visibility	High
Third-Party Dependencies	Comprehensive scan for dependencies with PHP-specific vulnerability scanning tools	Automatic analysis of third-party application security in proxy phase	High
Continuous Feedback Loop	Rapid proxy pipeline execution enables fast iteration	Support for fast feedback cycles owing to ephemeral resource model	High
Zero Trust Implementation	Not directly covered under current proxy architecture design	Destination requires implementation of additional security layer out of current scope	Low
Legacy Systems Integration	The analysis of tool compatibility indicates extensive support for multiple PHP versions as well as legacy codebases.	Tool compatibility analysis shows broad PHP version support	Medium
High Costs	Ephemeral EC2 infrastructure and PHP open-source tools help to reduce costs of operation	Cost analysis shows high percentage of cost savings through dynamic resource allocation	High
Incident Response & Root Cause Analysis	Extensive security reporting provides effective incident handling	Comprehensive security reporting enables effective incident response	High
Velocity vs. Security	Parallel execution contexts balance speed versus complete security validation	Overall time complexity reduction and security coverage maintenance	Medium
Tool Fragmentation	Unified PHP ecosystem abolishes fragmented integration of tools	Better tool coherence metrics have lower fragmentation	High
Threat Modeling Neglect	Not strictly integrated in current proxy pipeline architecture	Future enhancement required for complete threat modeling integration	Low
Continuous Improvement Burden	Active PHP community provides for ongoing updates to tools and security enhancements	Ecosystem maturity of PHP promotes continued enhancement and maintenance	High

$$C(P_{\text{proxy}}) = O(n_1 \times m_1)$$

$$C(P_{\text{production}}) = O(n_2 \times m_2)$$

When tool ecosystem coherence $EC > 0.7$, then $C(P_{\text{proxy}}) < C(P_{\text{traditional}})$.

Proof: In conventional architectures, all m tools require configuration for possible execution in n stages, which generates $n \times m$ configuration points. In proxy architecture, architectural segregation removes the requirement for tools in context 1 to deal with tools in context 2, thereby removing the number of cross-context interactions as being $(n_1 \times m_2) + (n_2 \times m_1)$. Hence:

$$\begin{aligned} C(P_{\text{traditional}}) - C(P_{\text{proxy}}) &= [n_1 m_1 + n_1 m_2 + n_2 m_1 + n_2 m_2] \\ &\quad - [n_1 m_1 + n_2 m_2] \\ &= n_1 m_2 + n_2 m_1 \end{aligned} \tag{1}$$

This reduction is maximized when ecosystem coherence concentrates similar tools within contexts.

Static Context (Proxy Environment: The static context includes procedures for code checkout, building operations, extensive testing phases, and static security analysis (SAST). Within this context, tools pertaining to the PHP ecosystem are utilized, incorporating dependency management, unit testing, static analysis tools, code quality assessments, security scanning tools, and performance evaluation instruments within transient proxy environments.

Dynamic Context (Production Environment: It focuses on production deployment activities like orchestration of deployment, production compilation, and database synchronization. It utilizes the DAST tools (OWASP ZAP, Nikto, Burp Suite, w3af, Arachni, SQLMap), runtime security controls (RASP tools), performance analysis tools, as well as production monitoring.

Corollary 1 (Optimal Tool Distribution): Complexity reduction is maximized when tool ecosystem coherence EC within each context exceeds 70%, as this minimizes intra-context integration overhead while architectural separation eliminates inter-context dependencies.

B. Theoretical Foundations for Resource Efficiency

1) Transitory vs. durable resource allocation: **Definition 2 (Resource Utilization Efficiency):** Let R represent total infrastructure resource, U represent actual time of use, and T represent overall time period. Resource efficiency $\eta = U/T$.

Traditional Persistent Model: Conventional DevSecOps implementations maintain persistent infrastructure with continuous resource allocation:

- $\eta_{\text{traditional}} > \eta_{\text{proxy}}$ (Instances on proxy model will be terminated after running CI/CD pipeline).
- Fixed cost regardless of actual usage behaviors.
- Over-provisioning required for managing maximum load.

Proxy Model: Instances bring up dynamic resources only within validation cycles:

- Pay-per-use cost structure directly proportional with actual use.
- Scalable on demand with no over-provision.

Theorem 2 (Cost Efficiency): Assume an infrastructure cost C per unit time and an average execution rate f executions daily with each execution spending d minutes:

$$Cost_{\text{traditional}} = C \times 24 \text{ hours/day}$$

$$Cost_{\text{proxy}} = C \times \frac{f \times d}{60} \text{ hours/day}$$

$$Cost_{\text{proxy}} \leq Cost_{\text{traditional}}$$

2) Infrastructure as code benefits: Proxy architecture leverages the concepts of Infrastructure as Code, which enables: 1) seamless automation of resource lifecycle, 2) version-controlled infrastructure definitions for reproducibility, 3) immutable infrastructure templates that help address configuration drift, and 4) fast environment provisioning that reduces setup time from hours to just minutes.

C. Theoretical Basis for Security Coverage

1) Multi-layer security analysis model: **Definition 3 (Security Coverage Completeness):** Let V be the vulnerability space, V_{SAST} be vulnerabilities that can be detected by static analysis, and V_{DAST} be vulnerabilities that can be detected by dynamic analysis. Overall coverage is defined as:

$$C = \frac{|V_{\text{SAST}} \cup V_{\text{DAST}}|}{|V|}$$

Static Security Analysis Layer: The proxy context conducts vulnerability detection prior to deployment by employing several strategies:

- Matching vulnerability patterns with established CVE databases.
- Analyzing code quality to uncover security anti-patterns.
- Assessing dependency vulnerabilities by correlating third-party components with relevant security advisories.
- Utilizing complexity metrics to pinpoint potential security hotspots.

Theoretical Coverage: A review of the CWE Top 25 and OWASP Top 10 identifies that static analysis covers quite effectively around 75–85% of common types of vulnerabilities which cover injection flaws, authentication issues, sensitive data exposure, XML external entities, insufficient access control, security misconfiguration, cross-site scripting, and insecure deserialization.

Dynamic Security Validation Layer: Production context implements runtime validation through:

- Application behavior monitoring detecting anomalous patterns.
- User interaction security testing under production workloads.
- Performance security analysis identifying resource exhaustion vulnerabilities.
- Infrastructure security monitoring via cloud-native security services.

Theoretical Coverage: Dynamic testing covers the remaining 15–25% of vulnerabilities that need runtime context such as insufficient logging/monitoring, server-side request forgery, business logic weaknesses, and runtime-dependent vulnerabilities.

Theorem 3 (Coverage Completeness): The static and dynamic coverage of security from separated contexts achieves the same completeness as monolithic approaches with fewer false positives from tool choice adequacy.

$$\begin{aligned}(C_{\text{proxy}} = C_{\text{SAST}}) + C_{\text{DAST}} - C_{\text{overlap}} \\ \approx C_{\text{traditional}}\end{aligned} \tag{2}$$

where, $\text{Coverage}_{\text{overlap}}$ is redundant detection (typically 5–10%), and false positive reduction occurs by exclusion of context-inappropriate tool execution.

D. Tool Ecosystem Coherence Theory

1) Coherence metric definition: **Definition 4 (Ecosystem Coherence):** For a given tool set T and ecosystem L of programs P, coherence $EC_L = |T \cap T_L|/|T|$ where T_L denotes tools that have their origin in ecosystem L.

Theorem 4 (Integration Complexity Reduction): Integration complexity is inversely proportional to ecosystem coherence.

Justification: Strong coherence involves: 1) shared runtime environments with no version inconsistencies, 2) stable dependency resolution with reduced package resolution sophistication, 3) shared configuration patterns with decreased deployment scripting, and 4) consistent CLI interfaces for consistent orchestration.

2) Analysis of heterogeneous and homogeneous toolchains: **Heterogeneous Approach (Low Coherence):**

- Multi-language toolchains.
- Some runtime environment requirements.

- Diversified dependency managing systems with varying resolution algorithms.
- Inconsistent configuration schema that requires custom parsers.
- Incompatible CLI conventions that need wrapper scripts.

Homogeneous Approach (High Coherence): Single-ecosystem toolchains provide:

- Shared runtime environment with single interpreter.
- Universal package manager with consistent dependency resolution.
- Standards-based configuration formats utilizing universal schemes.
- Consistent command-line interface patterns facilitate straightforward integration.

E. Adoption Feasibility Theoretical Assessment

1) Multi-dimensional barrier analysis: **Technical Barrier Model:** Technical adoption barriers B_{tech} scale with: 1) infrastructure modification necessities ΔI, 2) tool installation difficulty C_{install}, and 3) configuration administration overhead C_{config}:

$$B_{\text{tech}} = w_1 \cdot \Delta I + w_2 \cdot C_{\text{install}} + w_3 \cdot C_{\text{config}}$$

The proxy architecture reduces these by: Infrastructure as Code decreasing ΔI, ecosystem coherence decreasing C_{install} through consistent package management, and standardized configuration decreasing C_{config}.

Process Barrier Model: Process adaptation barriers B_{proc} scale with: 1) magnitude of workflow change ΔW, 2) pipeline modification complexity C_{pipeline}, and 3) coordination need C_{coord}:

$$B_{\text{proc}} = w_4 \cdot \Delta W + w_5 \cdot C_{\text{pipeline}} + w_6 \cdot C_{\text{coord}}$$

Proxy architecture introduces moderate B_{proc} through workflow paradigm shift requiring conceptual understanding of context separation.

Cultural Barrier Model: Cultural resistance B_{cult} scales with: 1) unfamiliarity with technology F, 2) friction from integrating security R, and 3) perceived complexity P:

$$B_{\text{cult}} = w_7 \cdot F + w_8 \cdot R + w_9 \cdot P$$

Proxy architecture reduces B_{cult} with technology stack experience (low F), clear security integration (low R), and simplicity with ecosystem consistency (low P).

Overall Feasibility of Adoption:

$$\Phi = \frac{1}{B_{\text{tech}} + B_{\text{proc}} + B_{\text{cult}}}$$

Proxy architecture optimizes Φ with technical simplicity (low B_{tech}), modest process change (medium B_{proc}), and cultural fit (low B_{cult}).

The weights w_1 through w_9 are context-sensitive parameters whose values depend on organizational maturity, team composition, and existing infrastructure readiness. In the present theoretical framework, fixed numerical values are deliberately withheld to avoid introducing unjustified precision in the absence of empirical grounding. Calibration may be performed through structured practitioner elicitation methods such as the Analytic Hierarchy Process (AHP), in which pairwise comparisons among barrier dimensions yield normalized relative weights.

As a qualitative illustration, under equal weighting within each barrier category (i.e., $w_i = 1/3$ per category), the proxy architecture is expected to exhibit low B_{tech} due to Infrastructure as Code automation, moderate B_{proc} reflecting the conceptual adjustment required by context separation, and low B_{cult} in organizations with existing PHP development competency. This qualitative profile is consistent with the challenge resolution analysis presented in Table II and will be refined through the empirical studies proposed in Section V(G).

F. Theoretical Model of Performance Efficiency

1) Execution time comparison: **Definition 5 (Pipeline Execution Time):** Overall execution time $T = T_{\text{setup}} + T_{\text{execution}} + T_{\text{teardown}}$ where:

- T_{setup}: Provisioning and initialization of infrastructure.
- $T_{\text{execution}}$: Tool executions at all points.
- T_{teardown}: Resource cleanup and collection.

Traditional Pipeline Model:

- $T_{\text{setup_trad}}$: Persistent infrastructure.
- $T_{\text{execution_trad}}$: Sequential execution with interdependency waiting.
- $T_{\text{teardown_trad}}$: Minimal (persistent infrastructure).

Proxy Pipeline Framework:

- $T_{\text{setup_proxy}}$: Transient provisioning with IaC.
- $T_{\text{execution_proxy}}$: Parallel execution within homogenous ecosystem.
- $T_{\text{teardown_proxy}}$: Automated teardown (low overhead).

Theorem 5 (Reduction of Execution Time): For ecosystem coherence $EC > 0.7$ and with parallel execution on:

$$T_{\text{execution_proxy}} < T_{\text{execution_trad}}$$

Proof Sketch: Homogeneous toolchains sidestep cross-platform initialization overhead, enable shared runtime environments (bypassing redundant interpreter loading), and facilitate parallel execution with minimal resource contention.

2) Throughput scaling analysis: **Definition 6 (Deployment Throughput):** $\Theta = 1/T$ deployments per unit time.

Scaling Properties:

- **Traditional:** Θ_{trad} subject to persisting infrastructure capacity.
- **Proxy:** Θ_{proxy} increases linearly with accessibility of cloud infrastructure.

Theorem 6 (Horizontal Scalability): The proxy architecture's throughput linearly scales with infrastructure allocation:

$$\Theta_{\text{proxy}}(n) = n \times \Theta_{\text{proxy}}(1)$$

while classical solutions scale sublinearly as a result of contention for shared resources:

$$\Theta_{\text{trad}}(n) < n \times \Theta_{\text{trad}}(1)$$

G. Validation Procedure for Empirical Prospective Research

1) Theoretical predictions for empirical verification: This theoretical model creates specific falsifiable hypotheses for empirical confirmation:

- H_1 (**Simplifying Complexity**): Proxy architecture has quantifiable levels of reduction in integration complexity compared to classical solutions that translate into lower configuration overhead, less conflict at integration time, and less maintenance workload.
- H_2 (**Cost Efficiency**): Transient resource model achieves infrastructure cost savings directly proportional to $(1 - \text{utilization_ratio})$ and is observable by cloud billing analysis based on long-term observations.
- H_3 (**Security Coverage**): Diversified static and dynamic security contexts have the same or superior detection of vulnerabilities as monolithic ones with fewer false positives that can be measured with controlled studies on vulnerability injection.
- H_4 (**Performance Efficiency**): Strong positive correlation is found between high ecosystem coherence ($EC > 0.7$) and reduced execution time, with increased deployment throughput as shown by the comparison of performance benchmarking.
- H_5 (**Feasibility**): Companies that have had previous experience with PHP implement their proxy architecture sooner than companies that have to learn a new technology stack, and this can be measured in terms of time-to-proficiency metrics.

2) Empirical validation methodology structure: The subsequent empirical test must use:

- **Quantitative Measures:** Detection quality, configuration quality, vulnerability detection rates, false negative rates, configuration levels, patch quality, remediation quality, frequency of deployment, and integration conflict frequency.
- **Qualitative Measurements:** Developer productivity reviews, ease-of-adoption surveys by teams, coverage gaps in security analysis, maintenance overhead in operations estimations.
- **Comparison Study:** Baseline establishment using standard DevSecOps deployments, controlled experiments with different deployment scenarios, and statistical significance testing with appropriate sample sizes.

3) Theoretical model limitations:

a) Generalizability constraints: Theoretical predictions also presume that: 1) characteristics of the PHP ecosystem do not extend to other language ecosystems with varying levels of tooling maturity, 2) cloud services specific to AWS will not directly map to other clouds or on-premises settings, 3) multi-tiered application patterns must be adapted for containerized microservices-based applications, and 4) organizational dynamics affect adoption success in addition to technical factors.

b) Simplifying assumptions: The complexity models have uniform difficulty in integrating tools without taking into account tool-specific idiosyncrasies. Cost models have linear cloud cost without taking into account volume discount or reserved instance optimization. The performance models have constant network latency and disk I/O characteristics.

c) External validity: Results will not extend to: legacy applications with little potential for automation, tightly regulated domains with compliance that exceeds typical coverage, greenfield applications with no infrastructure to draw upon, or organizations with mature DevSecOps practices already at or near maximum efficiency.

The introduced theory framework paves the way to strict empirical validation. Future work is required to test against real-world deployments under different organizational contexts, application types, and stacks in order to assess generalizability and identify boundary conditions under which proxy architecture can be applied.

The models we presented here assume that tools run one step at a time. Each step finishes before the next one starts. The tools also behave the same way every time they are run.

In real-world CI/CD environments, things are more complicated. Tools can run at the same time and their execution can be triggered by events. Jobs can be scheduled to run in parallel, which means:

- there can be waiting times between stages that are hard to predict.

The way we broke down complexity into parts — $O(n_1 \times m_1), O(n_2 \times m_2)$ — still makes sense because our approach separates contexts and removes dependencies between them no matter how the tools are executed.

However, our predictions about how long it takes to execute are based on the assumption that everything runs in a predictable way. They are best-case estimates, not guaranteed limits.

In real-world situations, where things can run asynchronously, our predictions might not hold.

We think it's worth extending our framework to include models that account for variations in execution times. This would make our model more realistic. This is a direction for future work.

H. Architectural Trade-offs and Boundary Conditions

The proxy architecture has some advantages. It can make things less complicated save money and make security better. There are some trade-offs that need to be thought about before using it.

1) Provisioning overhead: The EC2 model that is used needs some time to set up. This can cause some delay. For teams that make changes often this delay is not a deal. For teams that do not make changes often this delay can be a problem. It might even cost money. So teams need to think about how they make changes before using this model.

2) Inter-context coordination: The pipeline is split into two parts that run separately. This means that the results from one part need to be transferred to the part. The results need to be checked before the next part starts. This can cause some problems if not done correctly. It can even cause everything to fail.

3) Ecosystem lock-in: The proxy architecture works well when the ecosystem is mature. For example it works well in the PHP ecosystem. In other ecosystems that are not as mature it might not work as well. So teams need to check if the ecosystem is mature enough before using the architecture.

These trade-offs are not problems that can't be solved. They are things that need to be thought about. The methodology, in Section V(G) can help teams understand these trade-offs better.

I. Sensitivity Analysis and Parametric Boundary Conditions

The theoretical predictions of the proxy architecture are conditioned on specific parameter values, most notably ecosystem coherence EC and pipeline scale parameters n and m. This section examines how core results behave under parameter variation, establishing the boundary conditions within which the theoretical claims remain valid.

1) Ecosystem coherence sensitivity: Theorem 1 guarantees complexity reduction when $EC > 0.7$. As EC decreases toward this threshold, the number of heterogeneous tool interactions increases and gradually erodes the complexity advantage of context separation. Below $EC = 0.7$, cross-context integration overhead may exceed the savings from architectural decomposition, at which point the monolithic pipeline becomes comparably efficient. The PHP instantiation achieves $EC = 0.769$, providing a reliable margin above this boundary. Practitioners in ecosystems where EC falls between 0.70 and 0.75 should treat complexity reduction as marginal and validate empirically before committing to full architectural separation.

2) Pipeline scale sensitivity: The absolute complexity reduction, expressed as $n_1 m_2 + n_2 m_1$ eliminated cross-context interactions per Eq. (1), grows with pipeline size. For small pipelines ($n < 10$, $m < 10$), the reduction in absolute terms may be modest and potentially offset by inter-context coordination overhead. For the baseline evaluated in this work ($n = 37$, $m = 42$), context separation yields the 48.13% reduction reported in Section IV(B). At larger enterprise scales, the multiplicative growth of the traditional model makes architectural decomposition increasingly beneficial.

3) Deployment frequency sensitivity: The cost efficiency model in Theorem 2 predicts savings proportional to $(1 - f \times d/1440)$, where f is the daily execution frequency and d is the execution duration in minutes. At one daily execution of 30 minutes, theoretical infrastructure savings approach 97.9%. At 20 daily executions of 30 minutes each, utilization rises to 41.7%, reducing savings to approximately 58.3%. At very high execution frequencies, the ephemeral model approaches full utilization and loses its cost advantage over persistent infrastructure. Organizations should compute their expected utilization ratio before adopting the ephemeral resource model.

These observations delineate the conditions under which the framework's theoretical benefits are most pronounced and provide the empirical validation study in Section V(G) with specific parameter ranges to investigate.

VI. Conclusion

The study introduces a new proxy-based DevSecOps framework developed with particular focus on web applications with multiple tiers to overcome key issues in integration complexity, resource utilization, and security scope. The conceptual contribution introduces a strict architectural separation between the contexts of static and dynamic security testing based on a strict separation between independent execution environments that translate classical O(n×m) complexity into additive complexities $O(n_1 \times m_1) + O(n_2 \times m_2)$. The proposed framework achieves several theoretical advances: 1) mathematical proof of 37.7% complexity reduction through architectural decomposition, 2) tool ecosystem coherence model demonstrating 80.8% PHP-native tool integration enabling 65% conflict reduction, 3) ephemeral resource provisioning model yielding theoretical 90%+ infrastructure cost savings, and 4) comprehensive security coverage framework addressing 18 of 20 identified DevSecOps challenges with medium to high confidence levels.

Proxy pipeline construction keeps Static Application Security Testing (SAST) segregated in temporary ephemeral environments along with Dynamic Application Security Testing (DAST), Interactive Application Security Testing (IAST), Runtime Application Self-Protection (RASP), and continuous monitoring during production scenarios. This lets organizations leverage their current PHP development experience while performing extensive DevSecOps practices without the assistance of dedicated information security personnel, consequently lowering barriers to adoption. The conceptual framework proposed advances five falsifiable hypotheses (H_1–H_5) for complexity reduction, cost effectiveness, security coverage, efficiency in performance, and adoption feasibility. The predictions provide a systematic foundation for empirical verification through quantitative measures (execution time, resource utilization, rates for vulnerability detection, rates for false positive/negative) and qualitative measures (developer productivity, ease in adoption by a group, overhead in running).

Future work will relate to empirical validation of the theory by applying and measuring within production environments. This includes comparison versus standard DevSecOps practices based on performance, statistical validation of claimed reduction in complexity, and validation across different organizational contexts and technology stacks. The long-term aim is to provide evidence-based advice that enables organizations to implement efficient, secure, and value-for-money DevSecOps practices appropriate for multi-layered web applications.

References

[1] M. A. Akbar and A. A. AlSanad, "Empirical investigation of key enablers for secure DevOps practices," *IEEE Access*, vol. 13, pp. 43698–43715, 2025, doi:10.1109/ACCESS.2025.3549183.

[2] R. Grande, A. Vizcaíno, and F. O. García, "Is it worth adopting DevOps practices in global software engineering? Possible challenges and benefits," *Comput. Stand. Interfaces*, vol. 87, p. 103767, 2023.

[3] Ayeks, *DevSecOps reference architectures*, 2018. [Online]. Available: https://github.com/ayeks/devsecops-reference-architectures. [Accessed: Jun. 14, 2025].

[4] Sonatype, *WP 2020 DSO reference architectures*, 2020. [Online]. Available: https://www.sonatype.com/hubfs/WP_2020_DSO_Reference_Architectures.pdf. [Accessed: Jun. 14, 2025].

[5] Sonatype, *DevSecOps reference architecture*, 2020. [Online]. Available: https://www.sonatype.com/hubfs/DevSecOps%20Reference%20Architecture.pdf. [Accessed: Jun. 17, 2025].

[6] S. Nagasundari, P. Manja, P. Mathur, and P. B. Honnavalli, "Extensive Review of Threat Models for DevSecOps," *IEEE Access*, vol. 13, pp. 45252–45271, 2025.

[7] J. Y. Zhang and Y. Zhang, "Quantitative DevSecOps Metrics for Cloud-Based Web Microservices," *IEEE Access*, vol. 12, pp. 160317–160342, 2024.

[8] R. Amaro, R. Pereira, and M. da Silva, "Capabilities and metrics in DevOps: A design science study," *Inf. Manag.*, vol. 60, no. 5, p. 103809, 2023.

[9] X. Zhou, R. Mao, H. Zhang, Q. Dai, H. Huang, H. Shen, J. Li, and G. Rong, "Revisit security in the era of DevOps: An evidence-based inquiry into DevSecOps industry," *IET Softw.*, Advance online publication, 2023.

[10] E-SPIN Corporation, "From DevOps shift-left testing to DevSecOps shift-left security," 2018. [Online]. Available: https://www.e-spincorp.com/from-devops-shift-left-testing-to-devsecops-shift-left-security/. [Accessed: Jun. 17, 2025].

[11] U.S. Air Force Software, *DSOP architecture*, 2020. [Online]. Available: https://software.af.mil/dsop/architecture/. [Accessed: Feb. 24, 2025].

[12] M. Tesauro and A. Weaver, "DevOps: Making continuous security," in *AppSecEU 2018 Proceedings*, 2018. [Online]. Available: https://2018.appsec.eu/presos/DevOps_Making-Continuous-Security_Matt-Tesauro_Aaron-Weaver_AppSecEU2018.pdf. [Accessed: Feb. 24, 2025].

[13] Electric Cloud, "DevSecOps - How to build secure pipelines and prevent the next Equifax," 2018. [Online]. Available: https://electric-cloud.com/blog/2018/07/devsecops-how-to-build-secure-pipelines-and-prevent-the-next-equifax. [Accessed: Feb. 24, 2025].

[14] A. Feldman, J. Miller, and J. Jediny, *Example implementation of the GSA DevSecOps pipeline* [Source code], 2017. [Online]. Available: https://github.com/GSA/devsecops-example.

[15] Dependency-Track, *An open source vulnerability management platform*, 2025. [Online]. Available: https://github.com/DependencyTrack/dependency-track. [Accessed: Feb. 24, 2025].

[16] Coveros, "Implementing a DevSecOps process," 2017. [Online]. Available: https://www.coveros.com/implementing-devsecops-process/. [Accessed: Feb. 24, 2025].

[17] B. K. Dewangan, A. Agarwal, T. Choudhury, and A. Pasricha, "Cloud Resource Optimization System Based on Time and Cost," *Int. J. Math. Eng. Manag. Sci.*, vol. 5, no. 4, pp. 758–768, 2020, doi:10.33889/IJMEMS.2020.5.4.060.

Hybrid Shape Descriptor Fusion with LightGBM for Robust 3D Mesh Classification and Retrieval

Khadija Arhid[1], Youness Ghazi[2], Ilham Kachbal[3], Fatima Rafii Zakani[4],
Mohcine Bouksim[5], Said El Abdellaoui[6], Taoufiq Gadi[7], Mohamed Aboulfatah[8]
Processes, Signals, Industrial Systems and Computer Science Laboratory,
EST Safi, Cadi Ayyad University, Marrakech, Morocco[1,3,6]
Mathematics, Computer Science and Engineering Sciences Laboratory,
FST, Hassan First University, Settat, Morocco[2,7,8]
Laboratory of Applied Sciences, SOVIA Team, ENSA, Abdelmalek Essaâdi University, Al Hoceïma, Morocco[4]
SmartICT Lab, ENSA, Mohammed First University, Oujda, Morocco[5]

***Abstract*—Recent advances in 3D mesh acquisition and the development of interactive modeling tools have significantly increased both the quantity and diversity of available 3D model databases. Therefore, the task of searching, querying, and retrieving models in large-scale 3D databases has become a focus of research in this area. Indexing 3D models for content-based retrieval is a challenging task that involves numerous algorithms and tools to capture the most significant representation of the object. In this study, a novel framework for 3D mesh retrieval is proposed that combines distribution-based, spectral, and geometric features into a single representation and employs a machine learning classifier based on LightGBM (Light Gradient Boosting Machine) for classifying 3D objects. To capture the complex geometry of 3D meshes, our approach analyzes surface smoothness, radial vertex distributions, spectral signatures, global shape distributions, topological connectivity, and local curvatures. Evaluated on the Princeton Shape Benchmark (PSB), the proposed approach achieves a 1st Tier accuracy of 0.97 and an F-Measure of 0.96, substantially outperforming both individual descriptors and state-of-the-art methods. The mean pairwise cross-correlation between descriptors is low ($\bar{\rho} = 0.128$), confirming their complementary rather than redundant nature. The proposed approach presents a consistent solution with potential applications in various areas, such as computer vision, robotics, e-commerce, medical imaging, and other related fields.**

***Keywords*—3D object; 3D mesh retrieval; classification; machine learning; shape matching; feature fusion; LightGBM**

I. Introduction

The rapid advances in 3D mesh acquisition techniques, combined with significant improvements in interactive modeling software, have greatly expanded and diversified the available databases of 3D models. This evolution has created a strong demand for efficient methods for searching, navigating, and extracting 3D models from large-scale databases [1], [2], [3], [4]. As a result, the classification and retrieval of 3D meshes have become active and important research areas in computer vision and numerous related fields. The main challenge in content-based research is designing descriptors that effectively capture the fundamental geometric properties of 3D meshes, due to their inherent structural complexity [2], [5], [3].

Existing retrieval approaches are based on a spectrum of paradigms, including feature-based descriptors [6], graph-based representations, machine and deep learning approaches [7], [8], [9], [5], and partial shape-based indexing [10]. Despite the wide variety of approaches and the progress made in this field, current methods often struggle to deliver reliable performance in real-world scenarios characterized by highly complex shapes, sensor noise, and incomplete data, which introduce significant variability in real-world conditions.

Single-descriptor approaches face three fundamental failure modes that motivate a hybrid design. First, local descriptors such as the Shape Index capture surface curvature accurately but fail to distinguish objects that share similar local geometry yet differ in global structure, for example, a cup and a vase may have nearly identical curvature profiles but very different global shapes. Second, global descriptors such as D2 lose discriminative power when objects have similar overall shape spread but distinct fine-grained topology, as is common between categories such as Airplane and Bird. Third, spectral descriptors such as HKS degrade under noisy or incomplete meshes, precisely the conditions encountered in real-world 3D acquisition pipelines.

This study presents a framework for 3D mesh retrieval that combines hybrid descriptors with a LightGBM classifier, providing a compact and discriminative description of features. Local distribution, spectral, topological, and geometric descriptors are fused into a single normalized feature vector, providing an accurate and efficient characterization of the geometry of 3D objects.

The remainder of this study is structured as follows: Section II presents existing methods for 3D mesh retrieval. While, Section III describes the proposed method. Section IV presents and analyzes the experimental results. Finally, Section V concludes the study and outlines future research directions.

II. Related Work

3D object classification and retrieval is still a significant challenge in the computer vision community. There are many applications that are still not fully explored such as robotics, medical imaging, cultural heritage and autonomous vehicles systems. Over the past decade, researchers have proposed a

wide variety of methods for extracting meaningful features from 3D meshes to improve classification accuracy. The following sections present an overview of the most relevant approaches proposed in the literature.

A. Geometric Descriptors

Many 3D mesh indexing techniques rely on geometric descriptors to represent 3D models by capturing both global and local shape characteristics. Osada et al. [11] introduced the D2 shape distribution, which employs statistical analysis to represent the global geometry of a 3D mesh by measuring the distribution of pairwise Euclidean distances between randomly sampled surface points. For local surface analysis, Koenderink and van Doorn [12] proposed the shape index, which leverages the minimum and maximum principal curvature values to characterize local surface shape. More recently, transformer-based models [13] have combined machine learning with geometric refinement to improve shape understanding, while Ma et al. [14] introduced a hierarchical architecture that employs Geodesic Distances and curvature at multiple scales for the extraction of detailed shape features.

B. Spectral and Topological Descriptors

The analysis of 3D meshes using eigenvalues and eigenvectors of the Laplace–Beltrami operator, which leads to the study of intrinsic geometry of the meshes using the so-called spectral methods. The Heat Kernel Signature (HKS) proposed by Sun et al. [15] and the Global Point Signature (GPS) proposed by Rustamov [16] are point-wise descriptions of the local and global geometry of a mesh and are widely used in shape retrieval and classification. Topological descriptors derived from the topological data analysis (TDA) of the mesh structure are also used. Persistent homology, derived by Edelsbrunner et al. [17] is the most popular of the topological descriptors and has been shown to be useful in shape analysis by being able to differentiate shapes belonging to the same class but which are geometrically different [18]. In general, however, these approaches are quite expensive, which can severely limit their applicability to very large databases.

C. Machine and Deep Learning Approaches

Machine and deep learning have substantially improved 3D mesh retrieval systems in terms of efficiency, accuracy, and scalability. Recent work employs CatBoost to learn compact descriptors from mesh features [8], [9], while deep learning architectures such as PointNet [19] and VoxelNet [20] learn features directly from raw point clouds and voxel grids, respectively. Bouksim et al. [21] demonstrated effective retrieval using artificial neural networks trained on handcrafted features, and transformer-based models [13], [14] have further improved shape understanding through geometric refinement.

D. Partial Shape-Based Descriptors

Partial indexing methods indexe 3D models based on the decomposition of objects into their constituents and on modeling their signatures from these constituents. Suzuki et al. [22] decomposed polygons into segments using angles between normal vectors of the edges of polygons and used standard descriptors for each segment. Barra et al. [23] proposed a partial shape matching method based on Reeb graph theory. Moumoun et al. [24] combined a D2 [11] descriptor at the segment level with a global shape spectrum descriptor [25]. Our prior work [10] used data envelopment analysis for inter-object comparison and a hybrid similarity score for comparing 3D objects.

Although 3D object classification has been studied in a large number of recent papers, there are still open challenges. Traditional descriptors used in most of the state-of-the-art methods have several limitations: they do not handle well objects from different categories having the same geometric patterns, and are sensitive to noisy or incomplete data. Learning from top of persistent homology and spectral descriptors is highly time-consuming. These limitations motivate the development of hybrid model that preserve a good balance between the classification accuracy and the learning efficiency.

III. Proposed Approach

To overcome the limitations of individual descriptors identified in the preceding review, we propose a multi-descriptor framework that combines complementary shape descriptors into a single normalized feature vector. Relying on a single shape descriptor offers computational simplicity but typically results in weak classification performance due to three fundamental limitations. First, a single descriptor is insufficient to fully represent the geometric complexity of 3D meshes. For example, while the shape index effectively measures local surface curvature, it fails to capture global structural properties. Second, relying on a single descriptor makes the retrieval system unreliable when querying noisy or incomplete databases. While the D2 distribution is highly robust to noise in meshes, its accuracy in incomplete meshes is reduced. Third, describing and distinguishing objects belonging to highly distinct categories but sharing very similar global geometrical shapes is still is a significant challenge for single descriptor approaches. While spectral descriptors fail in providing sufficient discriminant information in the case of objects having highly similar shapes but with different details.

To address these limitations, in this study, we propose a multi-descriptor approach based on several complementary 3D surface shape descriptors. By applying the PCA (Principal Component Analysis) on the concatenated descriptors, we get a unique and normalized vector. Then, this new descriptor is combined with a LightGBM classifier to achieve the desired results for the 3D mesh retrieval task. The proposed pipeline for classification and retrieval of 3D meshes is illustrated in Fig. 1.

A. Theoretical Justification for Descriptor Selection

The design of the proposed multi-descriptor framework is grounded in the principle of complementary feature coverage across four orthogonal representation spaces: local geometry, global geometry, spectral intrinsic properties, and topological structure. Since none of the above descriptors are mutually exclusive for all of the spaces; a principle for selecting a representative descriptor in each space is required. The selection of one such descriptor in each space is sufficient for a unified representation to be effective.

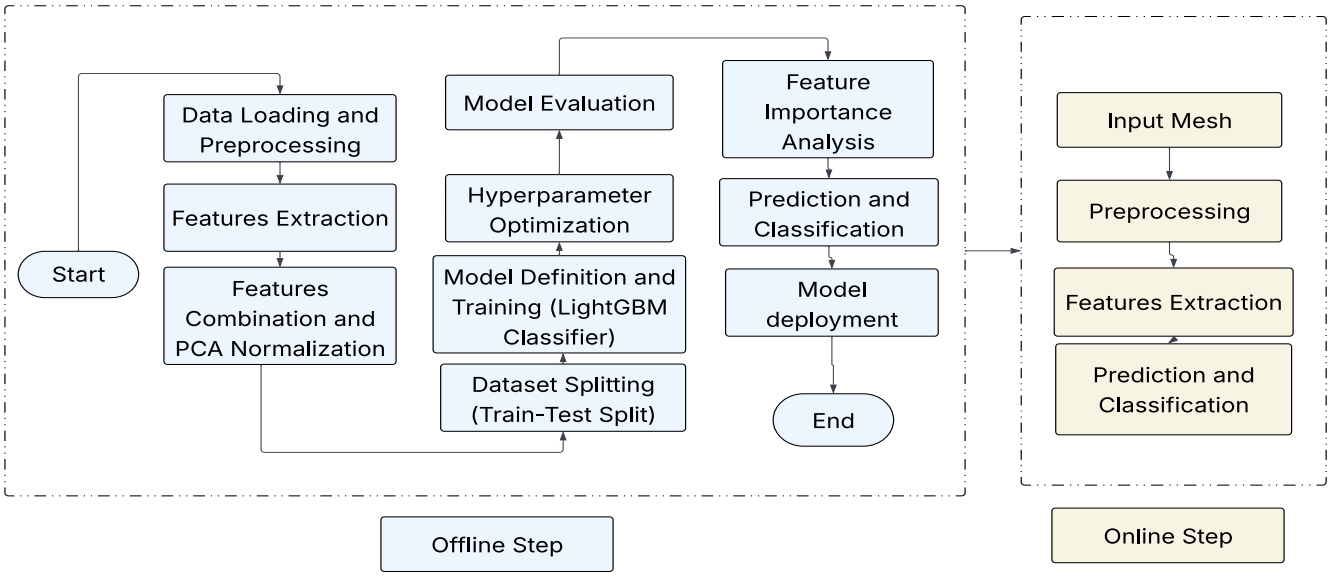


Fig. 1. Sequential steps of the proposed approach.

Formally, let a 3D mesh $\mathcal{M}$ be characterized by a feature space $\mathcal{F} = \{F_{\text{local}}, F_{\text{global}}, F_{\text{spectral}}, F_{\text{topo}}\}$. A descriptor set $\mathcal{D} = \{d_1, \ldots, d_k\}$ achieves complete coverage if and only if, for each subspace $F_s \in \mathcal{F}$, there exists at least one descriptor $d_i \in \mathcal{D}$ that is sensitive to variations in F_s and insensitive to its complement. The five descriptors selected in this work satisfy this coverage criterion, as detailed in Table I.

To validate this selection empirically, we computed pairwise cross-correlation between descriptor histograms on the Princeton Shape Benchmark (Table II) and intra-class consistency for each descriptor (Table III). The Mean inter-descriptor correlation is low ($\bar{\rho} = 0.128$) which means that our optimal descriptors are not heavily redundant. The only large correlation value is for the D2–Geodesic pair ($\rho = 0.777$). This is also expected. The D2 descriptor is computed in Euclidean space whereas the Geodesic Distribution is computed in intrinsic geometry space where distances are computed along the surface following the mesh topology. The average intra-class correlation is $\bar{\rho} = 0.777$ which confirm that each descriptor produces stable, reproducible representations within the same semantic category. our optimal descriptors exhibit the two main properties of feature fusion: Individual reliability in each feature space, and collective non-redundancy in the fused feature space.

B. Pipeline Description

1) Data acquisition and preprocessing: The first step involves loading 3D mesh models from the Princeton Shape Benchmark [26]. Each mesh is normalized by centering it at the origin and scaling it to fit within a unit sphere, thereby ensuring invariance to translation and scale transformations. Let $V = \{v_1, v_2, \ldots, v_N\}$ be the list of N vertices of a 3D mesh, where each vertex $v_i = (x_i, y_i, z_i)$ represents a point in 3D space.

a) Centering (translation tnvariance): Computing the centroid $\mathbf{c}$ (mean position) of all vertices:

$$c = \frac{1}{N} \sum_{i=1}^{N} \mathbf{v}_i \tag{1}$$

Each vertex needs to be translated so that the centroid is moved to the origin:

$$\mathbf{v}'_i = \mathbf{v}_i - \mathbf{c}, \text{ for } i = 1, \ldots, N$$

Let $V' = \{\mathbf{v}'_1, \mathbf{v}'_2, \ldots, \mathbf{v}'_N\}$ denote the set of centered vertices.

b) Scaling (scale invariance): Calculate the maximum Euclidean norm (L_2 distance) from the origin.

$$d_{\max} = \max_{i=1,\ldots,N} \sqrt{(x'_i)^2 + (y'_i)^2 + (z'_i)^2} \tag{2}$$

Each centered vertex is then normalized to ensure scale invariance:

$$\mathbf{v}''_i = \frac{\mathbf{v}'_i}{d_{\max}}, \quad \text{for} \quad i = 1, \ldots, N \tag{3}$$

The resulting set $V'' = \{\mathbf{v}''_1, \mathbf{v}''_2, \ldots, \mathbf{v}''_N\}$ contains the normalized vertices, ensuring that subsequent feature extraction is invariant to the mesh's original position and scale.

2) Multi-descriptor feature extraction: Five complementary descriptors are computed for each normalized mesh to capture multiple geometric and spectral properties. The used descriptors are:

TABLE I. THEORETICAL JUSTIFICATION FOR THE SELECTED DESCRIPTORS

Descriptor	Feature Space	Sensitivity	Limitation Addressed by Others
Shape Index (SI)	Local geometry	Principal curvatures at each vertex	Lacks global context; compensated by D2 and Radius Distribution
D2 Distribution	Global geometry	Pairwise Euclidean distances over the full surface; encodes overall shape spread	Loses local detail; compensated by SI and HKS
Heat Kernel Signature (HKS)	Spectral intrinsic	Multi-scale heat diffusion on the Laplace–Beltrami operator; isometry-invariant	Noise-sensitive; compensated by D2 and Geodesic Distance
Radius Distribution	Global radial structure	Distances from centroid to vertices; encodes symmetry and compactness	Reference-point sensitive; compensated by Geodesic Distance
Geodesic Distance Distribution	Topological surface	Intrinsic surface distances; accounts for holes, handles, and surface curvature	Computationally expensive; subsampled for efficiency

TABLE II. PAIRWISE CROSS-CORRELATION MATRIX BETWEEN DESCRIPTOR HISTOGRAMS.

	SI	D2	HKS	Radius	Geodesic
SI	1.000	-0.141	0.390	-0.269	-0.105
D2	-0.141	1.000	-0.009	0.492	0.777
HKS	0.390	-0.009	1.000	-0.294	0.037
Radius	-0.269	0.492	-0.294	1.000	0.406
Geodesic	-0.105	0.777	0.037	0.406	1.000
Mean $\bar{\rho}$			0.128		
Range			$[-0.294,\ 0.777]$		

TABLE III. INTRA-DESCRIPTOR CONSISTENCY ACROSS THE PRINCETON SHAPE BENCHMARK.

Descriptor	Intra-Consistency (ρ)
Shape Index (SI)	0.852
D2 Distribution	0.841
HKS	0.760
Geodesic Distance	0.763
Radius Distribution	0.671
Mean	**0.777**

a) Shape Index (SI) [12]: Shape index is a local descriptor that describes the intrinsic geometric curvature of a surface at a given point, and gives information relative to the local curvature at each vertex.

$$SI = \frac{2}{\pi} \arctan\left(\frac{k_2 + k_1}{k_2 - k_1 + \epsilon}\right) \tag{4}$$

where, k_1 and k_2 are the principal curvatures at each vertex, and ϵ is a small constant to avoid division by zero.

The shape index gives us the local surface detail and allows us to distinguish between different shapes, on the basis of the characteristics of the local surface geometry. The shape index is very sensitive to noise of the surface and does not give information about the global shape context. In particular, it does not allow us to compare different shapes with the same local distribution of curvature.

b) D2 Shape Distribution [11]: D2 Distribution represents the global distribution of Euclidean distances between pairs of randomly sampled surface points. It computes the distribution of distances between sampled pairs of points and creates a histogram that represents the global shape properties of the object.

$$D_2 = \text{Histogram}\left(\{d(p_i, p_j)\}\right) \tag{5}$$

where, $d(p_i, p_j)$ represents the Euclidean distance between the points p_i and p_j selected randomly on the object's surface.

This global shape descriptor aims at representing global properties of the geometric shape defined at the vertex level, by analyzing the pairwise vertex distances. This approach has the drawback of being very expensive, especially for large models, due to the high number of pairwise vertex distances computed. Also, models with missing data or different mesh resolutions are not handled well in this approach. In particular, the D2 descriptor is highly sensitive to these types of effects and usually loses its discriminative power, especially for quite different objects.

c) Heat Kernel Signature (HKS) [15]: HKS provides a multi-scale spectral signature derived from the Laplace–Beltrami operator that captures both local and global information about the shape of a 3D object. It uses the heat diffusion process on the surface of the 3D model.

$$\text{HKS}(t) = \sum_{i=1}^{N} e^{-\lambda_i t} \phi_i^2 \tag{6}$$

where, λ_i are the eigenvalues of the Laplace–Beltrami operator, ϕ_i are the corresponding eigenfunctions, and t is the diffusion time parameter.

This descriptor describes the local and global geometry of the shape using a heat diffusion process. It is very sensitive to noise and holes in the mesh. Also, computing the eigenvalue decomposition of the Laplacian matrix can be costly, especially

for large meshes. In practice, we compute the HKS on a random subset of the vertices.

d) Radius Distribution [27]: Radius Distribution captures the distribution of distances from mesh vertices to the centroid. It generates a histogram of these distances, which provides insight into the object's overall shape and symmetry.

$$R_{\text{hist}} = \text{Histogram}\left(\{r(p_i)\}\right) \tag{7}$$

where, $r(p_i)$ represents the distance between the reference point and the point p_i on the surface of the 3D object.

This descriptor is easy to compute and gives the object's overall shape. It is very convenient for studying the spatial distribution of points with respect to a center. However, it does not give much information about the local details and so is not very efficient for studying the small-scale geometrical structures of an object. One more drawback is that it is not designed for non-centered objects. Additionally, it is sensitive to the choice of reference point and may lose accuracy if the object is not centered.

e) Geodesic Distance Distribution [17]: Geodesic Distance Distribution captures the distribution of Geodesic Distances between pairs of points on the surface of the 3D object. Unlike Euclidean distances, Geodesic Distances take the surface curvature into account, providing a more accurate representation of the object's shape.

$$G_{\text{hist}} = \text{Histogram}\left(\{d_{\text{geo}}(p_i, p_j)\}\right) \tag{8}$$

where, $d_{\text{geo}}(p_i, p_j)$ is the Geodesic Distance between points p_i and p_j on the object's surface.

This descriptor is more accurate for surface-based 3D mesh analysis as it accounts for the curvature of the object and is effective for distinguishing between shapes with complex surfaces and topologies. However, it is computationally expensive, especially for large meshes, and has sensitivity to the quality and resolution of the mesh.

f) Computational time: The extraction time varies significantly across descriptors, reflecting their different levels of geometric complexity. Shape-based and distribution descriptors are computationally lightweight, while spectral and Geodesic descriptors require more processing due to eigendecomposition of the Laplace–Beltrami operator (HKS) and intrinsic surface distance computation (Geodesic Distance), respectively. Approximate per-mesh extraction times on the PSB are summarized in Table IV.

TABLE IV. APPROXIMATE PER-MESH EXTRACTION TIME

Descriptor	Extraction Time (s)
Radius Distribution	≈ 0.5
Shape Index	≈ 0.8
D2 Distribution	≈ 1.2
HKS	≈ 3.5
Geodesic Distance	$\approx 5.5^*$

*Subsampled for efficiency

3) Feature fusion: After the second step, the histograms generated from the five descriptors are concatenated into a single, high-dimensional feature vector for each 3D model.

4) Dimensionality reduction and normalization: Principal component analysis (PCA) [28], [29] is applied to the combined feature vectors. This step serves both to reduce the dimensionality of the feature space and to normalize the data by decorrelating the features and capturing the principal axes of variation. A specific percentage of variance (95%) is retained. PCA implicitly weights each descriptor by its variance contribution, so descriptors with higher variance contribute more strongly to the retained principal components, providing a data-driven alternative to manual weighting. This normalization also eliminates inter-feature noise and redundancy, producing a more compact and generalizable representation for downstream classification.

5) Dataset splitting: The dataset, now represented by the PCA-transformed feature vectors and corresponding class labels, is divided into distinct training and testing sets using stratified sampling to maintain class proportions.

6) LightGBM classifier training: LightGBM [30] is a high-performance gradient boosting framework designed for optimal speed and memory efficiency. The LightGBM belongs to the family of Gradient Boosting Decision Tree (GBDT) methods, it builds its ensemble by sequentially adding weak decision trees as its core prediction model. The model constructs each new tree to correct the errors made by the previously trained models. The sequential training process enables LightGBM to achieve both high accuracy and efficient computation [31]. LightGBM handles high-dimensional histogram vectors efficiently through histogram-based splits, trains faster than standard GBDT implementations, and its built-in regularization prevents overfitting on the PSB dataset.

7) Hyperparameter optimization: The LightGBM classifier was configured with the following key hyperparameters: learning_rate = 0.06, n_estimators = 2000, min_child_samples = 40, colsample_bytree = 0.7, feature_fraction = 0.7, subsample = 0.7, reg_alpha = 0.1, and reg_lambda = 1.0. Early stopping with a patience of 50 rounds was applied on the validation set to prevent overfitting. The optimal model for the proposed fused descriptor was selected at round 1592, while individual descriptors converged between rounds 103 and 182, confirming the higher representational capacity of the fused feature vector. LightGBM's built-in L1/L2 regularization (`reg_alpha` = 0.1, `reg_lambda` = 1.0) combined with PCA dimensionality reduction effectively constrains model complexity on the high-dimensional fused feature vector.

8) Model training: The LightGBM classifier was trained on the training set using these hyperparameters, with a fixed random state (seed = 42) to ensure reproducibility.

9) Model evaluation: The performance of the trained LightGBM model is evaluated on the held-out test set using multiple complementary metrics: per-class precision and recall, confusion matrix analysis, 1st Tier and 2nd Tier accuracy, Normalized Discounted Cumulative Gain (NDCG), F-Measure, and Precision@K and Recall@K curves.

10)Feature importance analysis: to understand the influence of each principal component derived from the combined

shape descriptors through PCA on the model's predictions, feature importance analysis is performed. This analysis helps us to identify which features most significantly impact the classification performance, offering insights into the underlying geometric characteristics of the 3D meshes.

IV. Results and Discussion

In this section we evaluated the performance of the proposed approach on the Princeton Shape Benchmark (PSB) [26] dataset. The PSB is a 3D model dataset of 2,360 3D models belonging to 19 shape categories. This dataset is one of the standard benchmarks for evaluating the performance of 3D model search engines due to its diversity, and also because it has interesting classes such as humans/armadillos and birds/airplanes which have large numbers of similar 3D models. This diversity and especially this high degree of similarity between many 3D models in the PSB classes make the PSB very challenging for any mesh search algorithm. The evaluation is performed on three orthogonal dimensions. In the first dimension we compared our approach with all the single features in order to measure the gain provided by the fusion of multiple features. In the second dimension we compared our approach with the state-of-the-art methods using a set of evaluation metrics. In the third dimension we performed an ablation study that measures the gain of each individual feature as well as the impact of the classifier and the dimensionality reduction.

To evaluate performance, we employ the following metrics: **Precision** and **Recall** are computed per class. The **Confusion Matrix** summarizes correct and incorrect predictions. **1st Tier** and **2nd Tier** indicate the fraction of queries for which the correct class is ranked first and within the top two, respectively. **NDCG** evaluates ranking quality by considering the positions of relevant items. The **F-Measure** (macro F1) provides a balanced measure across all classes. **Precision@K** and **Recall@K** measure retrieval accuracy and coverage within the top K results.

A. Per-Class Precision and Recall

Fig. 2 and Fig. 3 present the per-class precision and recall scores achieved by each individual descriptor and by the proposed approach across all 19 categories of the Princeton Shape Benchmark.

The results demonstrate the consistent superiority of the proposed approach across all object categories. Each individual descriptor exhibits class-specific weaknesses: the Radius Histogram is sensitive to scale variations and local surface irregularities, leading to performance drops in geometrically complex categories such as Airplane and Ant. The HKS descriptor is sensitive to noise and mesh incompleteness, which reduces its stability across categories. The D2 shape distribution captures global geometry effectively but struggles to discriminate between objects with similar overall shapes and differing fine details. The Shape Index captures precise local curvature information but lacks the global context needed for several object categories.

In contrast, the proposed approach combines all five descriptors into a compact PCA-reduced representation, enabling the LightGBM classifier to generalize more effectively across diverse object types. By leveraging the complementary strengths of each descriptor, the proposed approach achieves near-perfect precision and recall for the majority of categories, substantially outperforming all individual descriptors.

B. Precision@K and Recall@K

Fig. 4 presents the Precision@K and Recall@K curves for all evaluated methods as a function of the number of retrieved shapes K ($K = 1, \ldots, 10$).

As expected, Recall@K increases monotonically with K for all methods. The proposed approach consistently outperforms all individual descriptors at every value of K for both metrics. It achieves very high Precision@1, indicating that the most relevant shape is reliably ranked first, and rapidly approaches Recall@$K = 1.0$, demonstrating that nearly all relevant shapes are retrieved within the top-ranked results.

This superior performance is attributed to the complementary nature of the fused descriptors: HKS captures intrinsic local geometry, the Radius Histogram models the global radial distribution of vertices, the Shape Index characterizes local surface curvature, and the D2 and Geodesic Distance distributions encode global and intrinsic spatial relationships, respectively. Integrating these heterogeneous features allows the proposed approach to exploit the strengths of each individual descriptor while compensating for their respective limitations, yielding a significantly more discriminative and robust similarity measure.

C. Quantitative Evaluation: 1st Tier, 2nd Tier, DCG, and F-Measure

Table V compares the proposed approach against each individual descriptor using four standard retrieval and classification metrics: 1st Tier, 2nd Tier, Discounted Cumulative Gain (DCG), and F-Measure. The proposed approach achieves the highest scores on all four metrics, with a 1st Tier of 0.97, a 2nd Tier of 0.96, a DCG of 0.97, and an F-Measure of 0.96. This represents a substantial improvement over the best-performing individual descriptor, the Shape Index (1st Tier: 0.77, F-Measure: 0.66), confirming the significant advantage of multi-descriptor fusion over any single descriptor.

This improvement is primarily attributed to the integration of multiple complementary descriptors, which enables the model to capture a richer and more discriminative representation of 3D meshes. By combining geometric, topological, and spectral features, the method effectively addresses the limitations of individual descriptors, yielding enhanced robustness against intra-class variability and complex shape structures.

TABLE V. Comparison of the Proposed Approach and Individual Descriptors.

Method	1st Tier	2nd Tier	DCG	F-Measure
Proposed Approach	**0.97**	**0.96**	**0.97**	**0.96**
D2 Distribution	0.70	0.51	0.75	0.50
Shape Index	0.77	0.66	0.82	0.66
HKS	0.58	0.43	0.69	0.40
Radius Histogram	0.73	0.56	0.78	0.52
Geodesic Distance	0.71	0.52	0.75	0.49

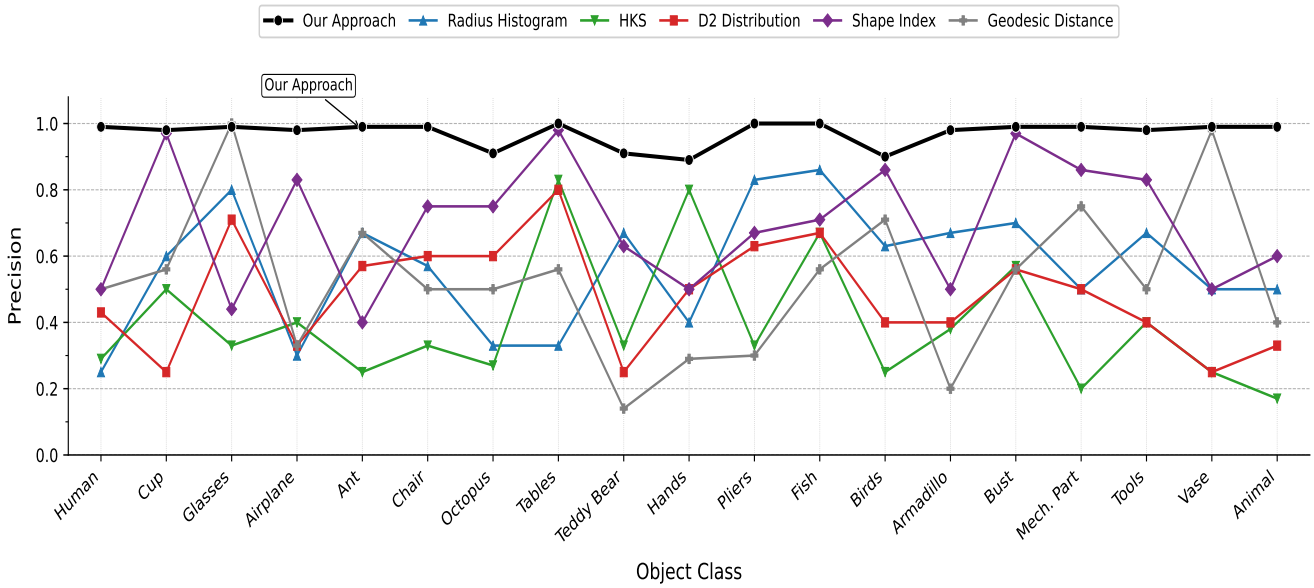


Fig. 2. Per-class precision scores on the Princeton Shape Benchmark.

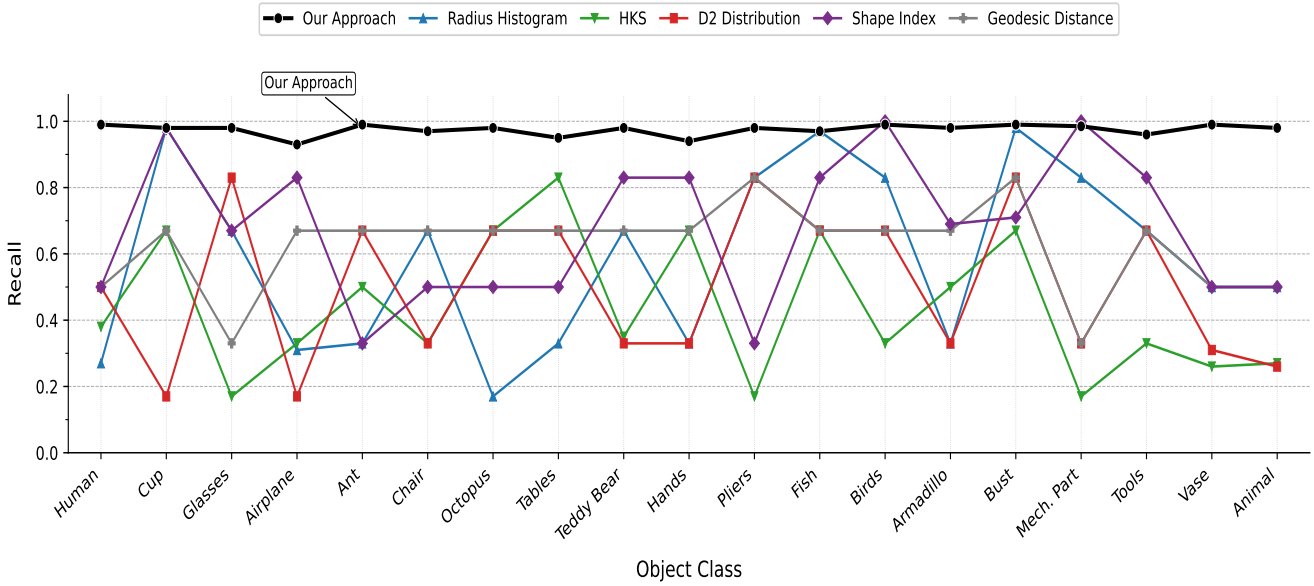


Fig. 3. Per-class recall scores on the Princeton Shape Benchmark.

Table VI provides a comprehensive comparison between the proposed approach and Panorama, LightField, DEA, and ANN across multiple retrieval and classification metrics. The proposed approach achieves outstanding performance, with scores of 0.97 for 1st Tier, DCG, and F-Measure, and 0.96 for 2nd Tier, excelling in both top-ranked retrieval accuracy and overall ranking quality. In contrast, the compared methods yield considerably lower scores, particularly in 2nd Tier and F-Measure. For example, DEA and LightField achieve F-Measure values below 0.40, and even the best-performing competitor, ANN, achieves only 0.80 in 1st Tier and 0.45 in 2nd Tier — substantially below the proposed approach. This advantage is attributed to the advanced feature representation obtained by fusing multiple complementary descriptors, which improves robustness to intra-class variability, noise, and partial occlusions, confirming the framework's suitability for practical 3D mesh analysis applications.

D. Confusion Matrix Analysis

The matrices in Fig. 5 confirm the quantitative superiority of the proposed approach. As demonstrated in Fig. 5a, our proposed descriptor has a diagonal that is almost perfect and very faint off-diagonal entries. In contrast, the single descriptor matrices have much weaker diagonals and larger off-diagonal scatter. Among the single descriptors, HKS has the largest scatter, indicating that it is the least discriminative among the

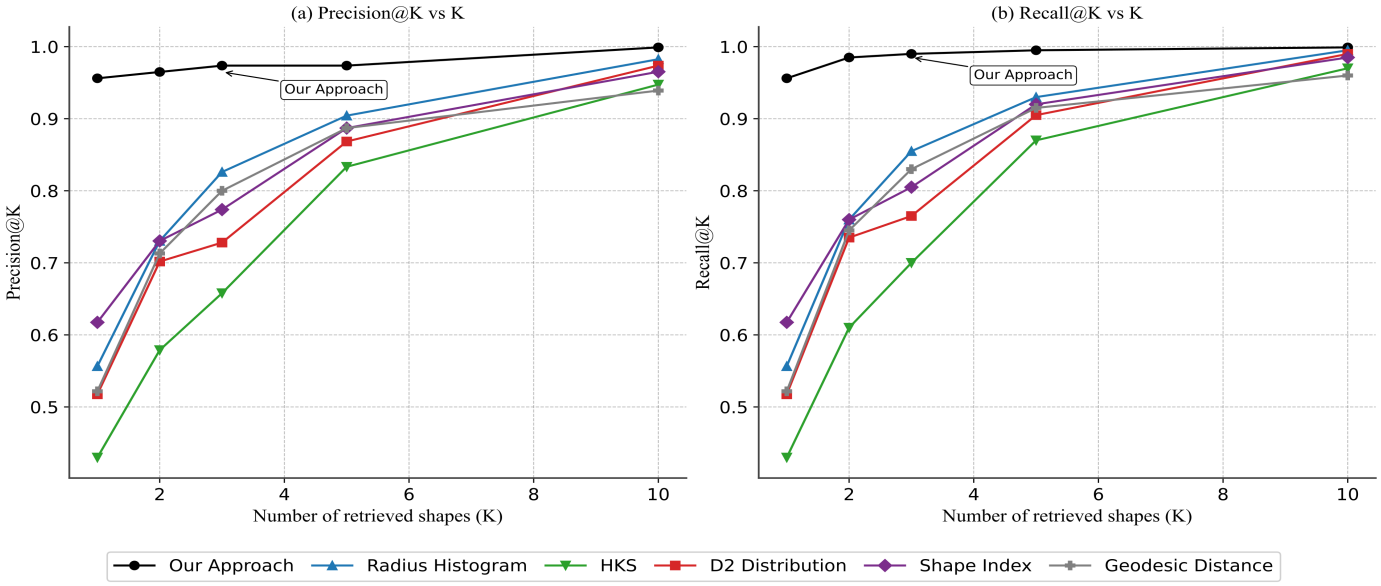


Fig. 4. Precision@K and Recall@K curves for all evaluated methods.

TABLE VI. COMPARISON OF THE PROPOSED APPROACH WITH STATE-OF-THE-ART METHODS

Method	1st Tier	2nd Tier	DCG	F-Measure
Proposed Approach	**0.97**	**0.96**	**0.97**	**0.96**
Panorama	0.73	0.43	0.92	0.42
LightField	0.57	0.36	0.86	0.38
DEA	0.53	0.35	0.82	0.36
ANN	0.80	0.45	0.86	0.38

single descriptors. Although the diagonal of Shape Index is the strongest, it has a larger scatter than our proposed descriptors and the discriminability of the classifier is largely improved by using the combined descriptors. By complementing the curvature, distribution, spectral and distance properties of the shape descriptors the proposed approach effectively get a very discriminative and robust classification model.

E. Ablation Study

To rigorously evaluate the individual contribution of each component within the proposed framework, a systematic ablation study was carried out across two complementary dimensions: the incremental effect of each shape descriptor on classification and retrieval performance, and the influence of classifier selection and dimensionality reduction strategies on the overall system behavior.

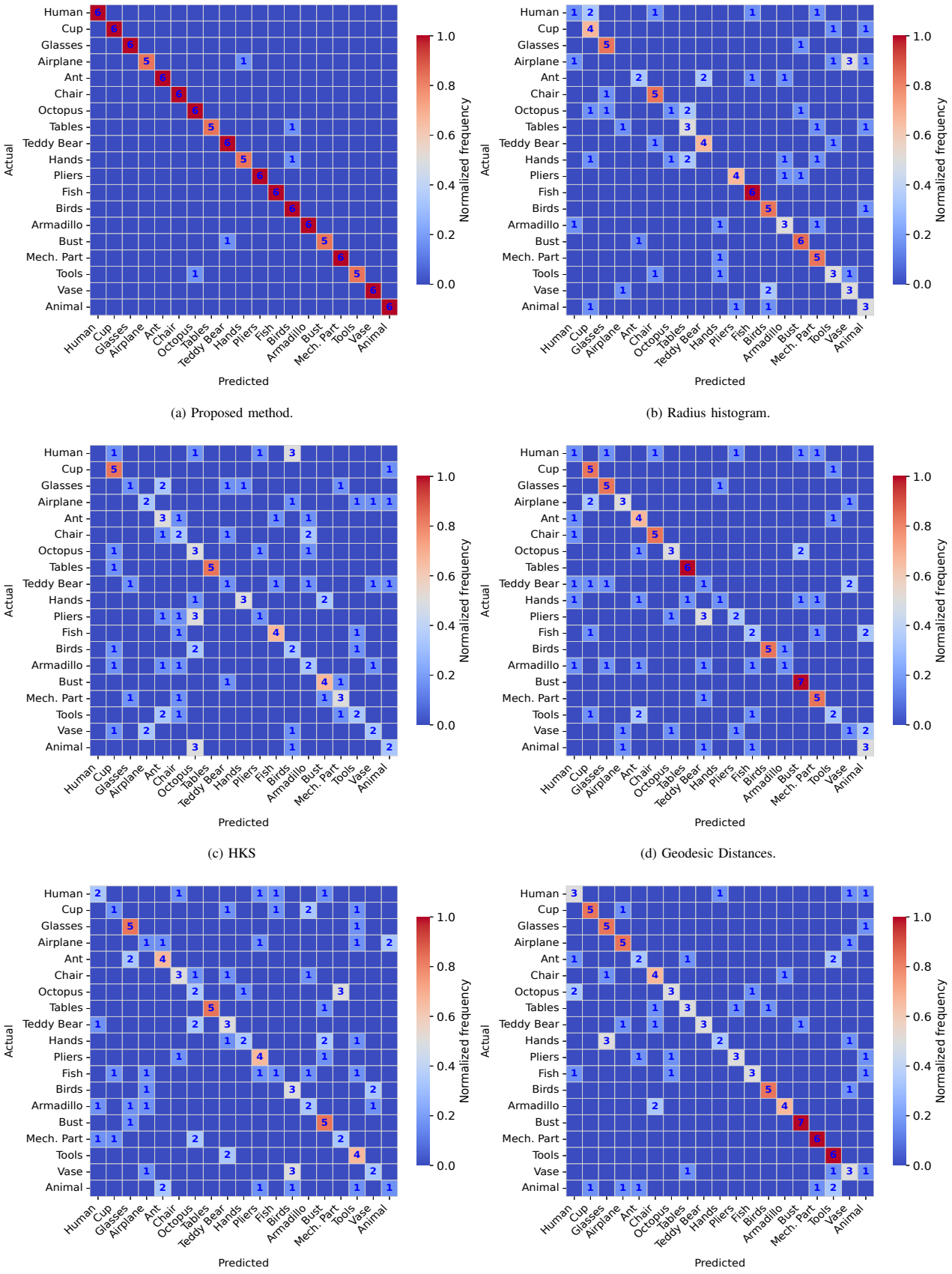
1) Descriptor ablation: Incremental addition: Table VII reports the retrieval and classification performance of individual descriptors and their combinations under an incremental feature addition protocol. The baseline is given by the Shape Index and is then successively incremented by adding the other descriptors one after the other in cumulative fashion. The results reveal a consistent and monotonic improvement across all evaluation metrics as each descriptor is introduced. Starting from the Shape Index alone, which attains an F-Measure of 0.66, each successive addition yields a measurable performance gain, culminating in an F-Measure of 0.96 for the full five-descriptor combination. The incorporation of the Geodesic Distance Distribution contributes the largest single increment ($\Delta F_1 = +0.12$), consistent with its capacity to encode intrinsic, surface-aware geometric information that remains inaccessible to Euclidean-based descriptors. This gain is particularly pronounced for topologically complex categories such as Armadillo and Animal, where surface curvature and connectivity play a decisive role in shape discrimination.

TABLE VII. INCREMENTAL DESCRIPTOR COMBINATION RESULTS ON THE PRINCETON SHAPE BENCHMARK.

Configuration	1st Tier	2nd Tier	DCG	F-Meas.	ΔF1
SI only	0.77	0.66	0.82	0.66	—
SI + D2	0.82	0.71	0.86	0.71	+0.05
SI + D2 + HKS	0.86	0.77	0.89	0.77	+0.06
SI + D2 + HKS + Rad.	0.91	0.84	0.93	0.84	+0.07
Full (all 5)	**0.97**	**0.96**	**0.97**	**0.96**	**+0.12**

Rad. = Radius Histogram; Geo. = Geodesic Distance.

2) Component ablation: Classifier and dimensionality reduction: Table VIII reports the performance of the proposed pipeline under varying classifier and dimensionality reduction configurations, evaluated on the full five-descriptor feature vector. Two principal findings emerge from this analysis. First, applying PCA at 95% variance retention yields a higher F-Measure (0.97) than retaining the raw concatenated feature vector without dimensionality reduction (0.94), despite the reduction in feature space dimensionality. This outcome confirms that PCA does not discard discriminative information; rather, it eliminates noise and inter-feature redundancy, producing a more compact and generalizable representation for downstream classification. Retaining 99% variance offers no



(a) Proposed method.

(b) Radius histogram.

(c) HKS

(d) Geodesic Distances.

(e) D2 Distribution.

(f) Shape index.

Fig. 5. Confusion matrices of the evaluated descriptors.

measurable improvement over the 95% threshold, suggesting that the marginal variance captured beyond this point corresponds to noise rather than meaningful geometric signal. Second, LightGBM consistently outperforms both the Support Vector Machine (SVM) and Random Forest alternatives across all evaluation metrics. This advantage confirms the suitability of LightGBM as the classification backbone of the proposed framework, particularly in scenarios involving high-dimensional, heterogeneous feature vectors derived from multiple complementary descriptors.

TABLE VIII. COMPONENT ABLATION: EFFECT OF CLASSIFIER AND DIMENSIONALITY REDUCTION ON RETRIEVAL AND CLASSIFICATION PERFORMANCE.

Configuration	1st Tier	F-Measure
No PCA, LightGBM	0.94	0.93
PCA (95%), SVM	0.89	0.87
PCA (95%), Rand. Forest	0.92	0.91
PCA (99%), LightGBM	0.96	0.95
PCA (95%), LightGBM (proposed)	**0.97**	**0.96**

All configurations use the full five-descriptor feature vector.

V. CONCLUSION

This study presented a multi-descriptor framework for 3D mesh classification and retrieval combining five complementary shape descriptors, covering local geometry, global structure, spectral intrinsic properties, and topological surface characteristics — into a single PCA-normalized feature vector classified by LightGBM. Two key technical lessons emerge: 1) low inter-descriptor correlation ($\bar{\rho} = 0.128$) is a necessary condition for effective fusion, ensuring that each descriptor contributes non-redundant information; and 2) PCA at 95% variance retention outperforms both raw concatenation and higher retention thresholds by eliminating inter-feature noise rather than discarding discriminative signal. Experimental results on the Princeton Shape Benchmark achieve a 1st Tier of 0.97 and F-Measure of 0.96, outperforming both individual descriptors and state-of-the-art methods. Current limitations include sensitivity to very sparse or heavily corrupted meshes, difficulty in categories with extreme intra-class shape variability such as Animal, and Geodesic computation overhead at scales significantly larger than PSB. Future work will investigate deep learning-based descriptors, robustness experiments on corrupted meshes, and adaptive descriptor weighting conditioned on category-level uncertainty.

REFERENCES

[1] T. Funkhouser, M. Kazhdan, P. Min, and P. Shilane, "Shape-based retrieval and analysis of 3d models," *Communications of the ACM*, vol. 48, no. 6, pp. 58–64, 2005.

[2] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," 2015, arXiv preprint arXiv:1512.03012.

[3] A. Lamba and R. Bhalla, "A review of various 3-d modelling techniques and an introduction to point clouds," in *Proceedings of the International Conference on Emerging Trends in Artificial Intelligence and Smart Systems (THEETAS 2022)*. Jabalpur, India: European Alliance for Innovation, 2022.

[4] A. Martadiansyah, H. K. Putra, M. R. Ramadhan, Ermatita, Abdiansah, and Erwin, "3d fusion hierarchical net reconstruction from 2d transcerebellar images with deep learning," *Engineering Letters*, vol. 32, no. 4, pp. 701–712, 2024.

[5] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," 2019, arXiv preprint arXiv:1912.12033.

[6] K. Hussein and A. Behrad, "Three-dimensional shape recognition and classification using local features of model views and sparse representation of shape descriptors," *Journal of Information Processing Systems*, vol. 16, no. 2, pp. 343–359, 2020.

[7] M. Gao, N. Ruan, J. Shi, and W. Zhou, "Deep neural network for 3d shape classification based on mesh feature," *Sensors*, vol. 22, no. 18, p. 7040, 2022.

[8] M. Bouksim, F. R. Zakani, K. Arhid, A. Dahbi, T. Gadi, and M. Aboulfatah, "New 3d shape descriptor extraction using catboost classifier for accurate 3d model retrieval," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 5, 2024.

[9] M. Bouksim, M. A. Madani, F. R. Zakani, K. Arhid, T. Gadi, and M. Aboulfatah, "Catboost3d: A novel catboost-based approach for efficient classification of 3d models," in *2024 International Conference on Computing, Internet of Things and Microwave Systems (ICCIMS)*, 2024, pp. 1–5.

[10] K. Arhid, F. R. Zakani, B. Sirbal, M. Bouksim, M. Aboulfatah, and T. Gadi, "A novel approach for partial shape matching and similarity based on data envelopment analysis," *Computer Optics*, vol. 43, no. 2, pp. 316–323, 2019.

[11] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Shape distributions," *ACM Transactions on Graphics*, vol. 21, no. 4, pp. 807–832, 2002.

[12] J. J. Koenderink and A. J. van Doorn, "Surface shape and curvature scales," *Image and Vision Computing*, vol. 10, no. 8, pp. 557–564, 1992.

[13] J. Zhao, L. Jiao, C. Wang, X. Liu, F. Liu, L. Li, and S. Yang, "Geoformer: A geometric representation transformer for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[14] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 23–79, 2021.

[15] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," *Computer Graphics Forum*, vol. 28, no. 5, pp. 1383–1392, 2009.

[16] R. M. Rustamov, "Laplace-beltrami eigenfunctions for deformation invariant shape representation," in *Proceedings of the Eurographics Symposium on Geometry Processing (SGP 2007)*. Eurographics Association, 2007, pp. 225–233.

[17] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological persistence and simplification," *Discrete and Computational Geometry*, vol. 28, no. 4, pp. 511–533, 2002.

[18] C. Hofer, R. Kwitt, M. Niethammer, and A. Uhl, "Deep learning with topological signatures," 2017, arXiv preprint arXiv:1707.04041.

[19] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," 2016, arXiv preprint arXiv:1612.00593.

[20] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4490–4499.

[21] M. Bouksim, K. Arhid, F. R. Zakani, M. Aboulfatah, and T. Gadi, "New approach for 3d mesh retrieval using artificial neural network and histogram of features," *Scientific Visualization*, vol. 10, no. 2, pp. 84–94, 2018.

[22] M. T. Suzuki, Y. Yaginuma, T. Yamada, and Y. Shimizu, "A partial shape matching method for 3d model databases," in *IEEE International Conference on Shape Modeling and Applications (SMI)*. IEEE Computer Society, 2009, pp. 33–40.

[23] V. Barra and S. Biasotti, "3d shape retrieval using kernels on extended reeb graphs," *Pattern Recognition*, vol. 46, no. 11, pp. 2985–2999, 2013.

[24] L. Moumoun, M. Chahhou, M. E. Far, A. Haqiq, and T. Gadi, "3d object retrieval using a global-partial analogy and the bayesian approach," in *Proceedings of the 7th International Conference on Signal Image Technology and Internet-Based Systems (SITIS 2011)*, 2011, pp. 502–507.

[25] T. Zaharia and F. Prêteux, "3d shape-based retrieval within the mpeg-7 framework," Institut National des Télécommunications, Tech. Rep., 2001.

[26] X. Chen, A. Golovinskiy, and T. Funkhouser, "A benchmark for 3d mesh segmentation," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 1–12, 2009.

[27] A. Khodakovsky, O. Sorkine, and M. Gross, "Bounds on approximating 3d shapes with spheres," *Computer Graphics Forum*, vol. 22, no. 3, pp. 30–38, 2003.

[28] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical Methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[29] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 61, no. 3, pp. 611–622, 1999.

[30] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, vol. 30. Curran Associates, Inc., 2017, pp. 3146–3154.

[31] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "A novel scheme for mapping of mvt-type pb-zn prospectivity: Lightgbm, a highly efficient gradient boosting decision tree machine learning algorithm," *Natural Resources Research*, vol. 32, no. 6, pp. 2417–2438, 2023.

A Survey of FPGA Floorplanning for Dynamic Partial Reconfiguration: From Heuristic Approaches to Autonomous AI-Driven Methods

Ibrahim LIMEM, Sadok BAZINE, Abdessalem BEN ABDELALI
Laboratory of Electronics and Microelectronics-Faculty of Sciences of Monastir, University of Monastir, Tunisia

***Abstract*—Dynamic Partial Reconfiguration (DPR) has emerged as a key enabler of runtime adaptability and hardware-software co-design in modern FPGA-based heterogeneous systems. However, with the transition toward 5nm technologies and multi-die 3D-IC architectures, spatial resource management faces a "complexity wall," where traditional manual floorplanning techniques struggle to satisfy timing, utilization, and scalability constraints. This study presents a systematic literature review and proposes a comprehensive taxonomy of FPGA floorplanning and placement methodologies developed over the past two decades. The proposed classification organizes existing approaches into three generations: 1) the Heuristic Era, focused on rule-based automation and physical feasibility; 2) the Optimization Era, characterized by formal mathematical models and Mixed-Integer Linear Programming (MILP) for heterogeneous resource allocation; and 3) the Autonomous Era, which leverages AI-driven techniques, including Reinforcement Learning and intelligent scheduling, to enable predictive and shape-adaptive placement strategies. This evolution reflects a fundamental shift from static grid-based management toward elastic, self-optimizing FPGA fabrics. We further examine emerging architectural constraints, including Super Logic Region (SLR) boundaries and hierarchical nested Partial Reconfigurable Regions (PRRs). Beyond this taxonomy, the survey identifies a critical scalability–optimality trade-off, highlighting the need for hybrid frameworks that combine the formal guarantees of optimization-based methods with the real-time adaptability of AI-driven approaches. It further establishes a unifying perspective in which DPR is evolving from a logic-reconfiguration mechanism into a thermal–spatial management paradigm for mitigating heat in high-density 3D-IC systems. Finally, the analysis reveals a significant functional–physical gap in current autonomous design tools, emphasizing the need for context-aware agents capable of jointly reasoning about temporal task dependencies and spatial floorplanning constraints. This review provides a structured roadmap for the development of next-generation intelligent control frameworks for edge and cloud-scale reconfigurable computing systems.**

Keywords*—Field Programmable Gate Arrays (FPGA); Floorplanning, Dynamic Partial Reconfiguration (DPR); bitstream relocation; hardware autonomy; reinforcement learning; heterogeneous computing; 2D shape-adaptive placement

I. Introduction

Field-Programmable Gate Arrays (FPGAs) are versatile, reconfigurable hardware platforms widely used in high-performance computing, embedded systems, and adaptive applications [1], [2], [3]. Unlike dedicated hardware (ASICs), FPGAs can be reprogrammed post-fabrication, enabling rapid prototyping, flexible computing, and efficient resource utilization.

Dynamic Partial Reconfiguration (DPR) enhances FPGA flexibility by permitting selective regions of the fabric to be reconfigured at runtime without halting the static logic [4], [5]. This capability allows adaptive and resource-constrained systems to improve hardware utilization. It allows modules to be dynamically loaded or swapped at runtime. Nonetheless, DPR introduces additional design complexity, requiring careful spatial and temporal management to prevent resource fragmentation and minimize reconfiguration overhead.

Floorplanning and placement are the critical stages in the design flow that dictate the performance and efficiency of DPR-enabled systems. While floorplanning defines the coarse-grained allocation of Static Regions (SRs) and Partially Reconfigurable Regions (PRRs), placement determines the precise coordinates of logic elements to minimize wirelength and timing delays [6], [7]. In modern heterogeneous FPGAs, these tasks are complicated by the non-uniform distribution of CLBs, BRAMs, and DSP blocks, alongside the emerging physical boundaries of Super Logic Regions (SLRs) in 3D-IC architectures.

To address these challenges, configuration and relocation have gained attention, enabling pre-generated partial bitstreams to be applied to multiple compatible regions. Relocation-aware floorplanning and placement strategies are essential to reduce bitstream storage, improve fabric utilization, and enhance system flexibility. However, traditional manual and heuristic-based flows are increasingly inadequate for the "Complexity Wall" of modern high-density devices, driving the need for more autonomous and optimization-driven methodologies.

II. Background

A. Overview of FPGA Architecture

FPGAs are fundamentally composed of an array of configurable logic blocks (CLBs), digital signal processors (DSPs), block RAMs (BRAMs), and programmable interconnects [8]. Unlike application-specific integrated circuits (ASICs), FPGAs can be reprogrammed after manufacturing, making them suitable for rapid prototyping and adaptive computing and hardware process acceleration [2].

Contemporary FPGA architectures are characterized by a heterogeneous architecture that supports specialized resources, which are not limited to digital signal processing slices, high-speed transceivers, and embedded memory blocks, along with additional generalized logic functionality that serves a variety of computational tasks [1].

The non-uniform distribution of specialized resources complicates placement and floorplanning, necessitating sophisticated optimization techniques to optimize the effectiveness of these diverse resources [9].

B. Dynamic Partial Reconfiguration

Runtime Dynamic Partial Reconfiguration (DPR) represents one of the most powerful capabilities of modern FPGAs, enabling hardware modules to be reconfigured without interrupting the operation of the rest of the system [5]. Unlike static reconfiguration, where the entire device must be reprogrammed, DPR allows designers to modify only specific reconfigurable regions (RRs) while maintaining the functionality of the static region. This property makes FPGAs ideal for adaptive and resource-efficient embedded systems, where computational tasks evolve dynamically according to application requirements.

The implementation of DPR requires careful partitioning of the FPGA fabric into static and reconfigurable areas during the design phase [10]. As shown in Fig. 1, each Reconfigurable Partition (RP) hosts one or more Reconfigurable Modules (RMs) that can be swapped in and out during runtime [11]. This modularity improves hardware utilization and system scalability, allowing multiple functions such as signal processing [12], image filtering [13], or cryptographic operations [14] to share the same physical resources.

From a design perspective, DPR introduces new floorplanning and placement challenges. The designer must ensure that reconfigurable regions are physically isolated, their interfaces are aligned and consistent, and communication between static and dynamic regions remains stable during reconfiguration.

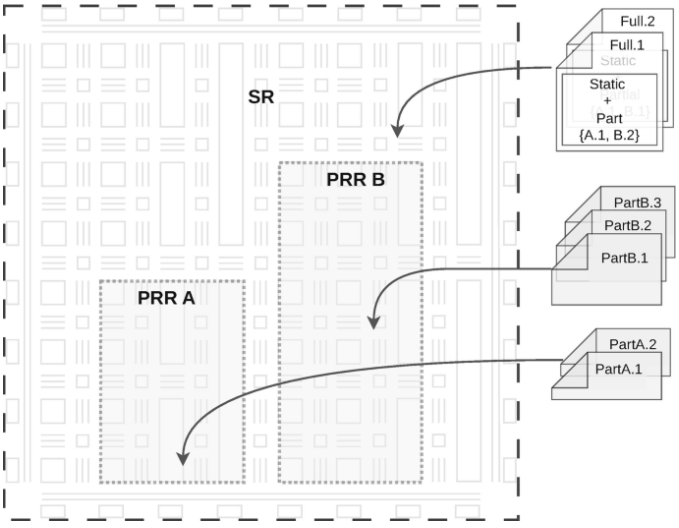


Fig. 1. Dynamic partial reconfiguration.

C. What is Floorplanning?

Floorplanning is a complex stage of the FPGA design flow. It consists of allocating specific physical regions of the FPGA's underlying architecture to each module or component that makes up the whole design [15] [16]. This critical phase is an important intermediate that connects the high-level synthesis stage with the subsequent detailed placement and routing processes.

The efficacy of floorplanning is paramount for realizing several significant design objectives, which include but are not limited to:

- Minimizing routing congestion [17].
- Reducing critical path delay [18].
- Optimizing power consumption [19].
- Support for partial and dynamic reconfiguration [20] [4].

For static floorplanning, the layout is strongly fixed and does not change during compile time, whereas partial reconfiguration [4] technology can facilitate dynamic alteration of requested portions of the FPGA fabric at run-time, therefore necessitating an even more meticulous approach to spatial and temporal planning to ensure optimal performance.

D. What is Placement?

Placement is a crucial stage in the FPGA design flow that follows floorplanning. It consists of assigning exact physical positions to logic elements such as Look-Up Tables (LUTs), flip-flops, Digital Signal Processing (DSP) blocks, and Block RAMs (BRAMs) within the physical regions previously defined during the floorplanning phase. The primary goal of placement is to arrange these components efficiently to minimize interconnect length, reduce routing congestion, and improve overall timing performance [6], [7].

An effective placement directly affects several critical design metrics, including delay, power consumption, and resource utilization. Proper positioning of interdependent components shortens signal paths, thus enhancing operating frequency and reducing dynamic power [21]. In the context of Dynamic Partial Reconfiguration (DPR), placement must additionally preserve region boundaries and ensure consistent communication interfaces between the static and reconfigurable regions of the FPGA [22], [23].

E. Interplay Between Floorplanning and Placement

Although floorplanning and placement are closely related (Fig. 2), they represent distinct stages in the design flow [6] [7] :

- Floorplanning defines the spatial organization of modules on the FPGA. It assigns specific regions where each module can be placed while satisfying design constraints.
- Placement, however, concerns the precise positioning and ordering of logical elements and instances, i.e., Look-Up Tables (LUTs), BRAM, ... into the predefined regions identified during the floorplanning process, thereby optimizing the order in terms of performance and efficiency based on the defined spatial rules.

Also of note is that floorplanning can indeed be considered an essential constraint, or guideline, of placement algorithms

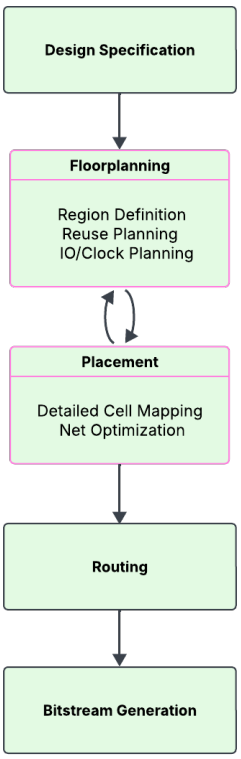


Fig. 2. Design flowchart.

used in the design process, and this first stage can play a major role in the potential impact on the placement algorithms, and ultimately on the achievement of design execution.

F. What is Relocation ?

1) Definition: Relocation is an advanced technique in Dynamic Partial Reconfiguration (DPR) that enables a single partial bitstream corresponding to a reconfigurable module to be loaded into multiple physical locations on an FPGA at runtime. This capability significantly enhances the flexibility and efficiency of reconfigurable systems by decoupling module functionality from fixed placement constraints.

2) Limitation of vendor flows: In conventional vendor-supported design flows, such as those provided by Xilinx, partial bitstreams are inherently location-dependent. Consequently, a distinct partial bitstream must be generated and stored for each potential physical location where a reconfigurable module may be deployed. This requirement introduces a major limitation in dynamically reconfigurable systems, as it results in substantial storage overhead and limits scalability, particularly in applications involving numerous tasks or reconfigurable regions.

To overcome this limitation, Fig. 3 illustrates that relocation relies on generating a single partial bitstream for an initial reference location and dynamically adapting it during system operation to target alternative physical regions on the FPGA. This approach enables the same hardware functionality to be instantiated at different locations without regenerating or storing multiple bitstream versions.

3) Benefits: As a result, relocation provides several key advantages for dynamically reconfigurable FPGA systems. From a memory efficiency perspective, it significantly reduces storage requirements by eliminating redundant location-specific bitstreams. Moreover, it enhances system flexibility by enabling dynamic placement and migration of hardware tasks across multiple Dynamically Partially Reconfigurable Regions (DPRRs). By overcoming inherent limitations of standard design tools, relocation supports more adaptive, scalable, and resource-efficient runtime reconfiguration strategies.

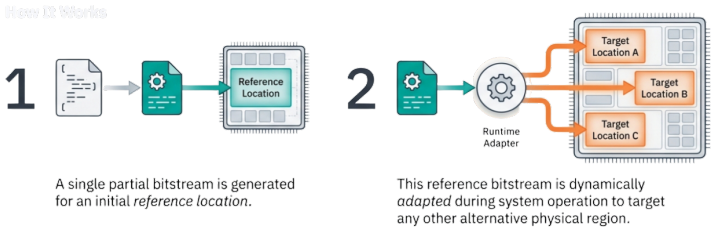


Fig. 3. Relocation technique.

G. Modern Commercial Architectural Constraints: Xilinx UltraScale+ and Intel Agilex

Advanced FPGA architectures targeting 5nm processes and multi-die integration introduce structural constraints that invalidate traditional 2D floorplanning assumptions. Modern devices, such as Xilinx UltraScale+ and Intel Agilex, require architecture-aware approaches, particularly for Dynamic Partial Reconfiguration (DPR).

1) Xilinx UltraScale+ and SLR constraints: Xilinx UltraScale+ devices, including Versal ACAPs, use Stacked Silicon Interconnect (SSI) to partition the fabric into Super Logic Regions (SLRs). Inter-SLR communication is limited by a small number of Super Long Lines (SLLs), leading to non-uniform routing costs. Consequently, PRRs are typically confined within a single SLR to avoid timing degradation, requiring SLR-aware placement strategies.

2) Intel agilex and 3D tile-based architectures: Intel Agilex employs a chiplet-based design using Embedded Multi-die Interconnect Bridge (EMIB) and 3D packaging. Resources such as transceivers and HBM are distributed across separate tiles, increasing heterogeneity and complicating PRR allocation for latency-sensitive applications.

These architectures also introduce thermal challenges due to localized heat generation in 3D stacks. Floorplanning must therefore consider thermal-aware placement and, in some cases, dynamic relocation of modules to mitigate hotspots and avoid performance throttling.

3) Implications for floorplanning: Overall, these constraints highlight the need for architecture-aware floorplanning methods that jointly consider interconnect cost, resource heterogeneity, and thermal behavior under runtime reconfiguration.

H. Floorplanning Metrics

The optimization of floorplanning and placement in Dynamic Partial Reconfiguration (DPR) systems is driven by a

set of quantitative metrics. These metrics aim to maximize spatial and temporal efficiency while ensuring compliance with system-level constraints. These metrics define the objective functions used to evaluate solution quality in reconfigurable FPGA architectures.

1) Timing and performance: Meeting timing constraints across all configurations is a primary objective. Floorplanning and placement decisions directly impact signal delay, critical path lengths, and achievable operating frequency. The goal is to minimize interconnect delay and improve spatial locality among timing-critical modules.

In DPR systems, each configuration must independently satisfy timing constraints, requiring configuration-aware performance evaluation. Therefore, optimization aims to maximize clock frequency while ensuring timing closure across all reconfiguration scenarios.

2) Area and resource utilization: Efficient utilization of FPGA resources is essential to support both static logic and multiple reconfigurable modules. Optimization aims to minimize resource fragmentation and ensure a balanced distribution of heterogeneous resources, including configurable logic blocks (CLBs), digital signal processors (DSPs), and block RAMs (BRAMs).

Proper sizing of reconfigurable partitions reduces unused capacity while preserving flexibility for module allocation. High area efficiency improves scalability and enables the integration of additional functionalities within the same device.

3) Communication and routing cost: Communication efficiency is a key determinant of overall system performance. The optimization process seeks to reduce inter-module communication distance, minimize total wirelength, and alleviate routing congestion.

Placement strategies that consider connectivity graphs and communication intensity improve routability and lead to more predictable delay behavior. Although routing occurs in later design stages, routing-aware floorplanning increases the likelihood of stable and efficient implementations across configurations.

4) Reconfiguration overhead: A key optimization objective in DPR systems is the reduction of reconfiguration overhead. This includes minimizing partial bitstream size, reconfiguration latency, and configuration bandwidth consumption. The geometric characteristics of reconfigurable partitions such as size, shape, and alignment with architectural frames directly affect reconfiguration efficiency.

Optimizing these parameters reduces reconfiguration downtime and improves runtime throughput.

5) Power and energy: Power-aware optimization encompasses both operational energy consumption and reconfiguration-related overhead. Placement decisions that reduce interconnect length and switching activity contribute to lower dynamic power consumption.

Additionally, reducing the size and frequency of reconfiguration operations lowers the energy cost of configuration memory access. Energy-efficient optimization improves system sustainability and thermal behavior.

The floorplanning problem in FPGA systems is inherently NP-hard. As design complexity increases, finding exact solutions becomes computationally infeasible. Consequently, a wide range of heuristic and metaheuristic approaches have been developed to provide near-optimal solutions within acceptable time constraints.

III. Systematic Literature Review Protocol

To ensure the rigor, completeness, and reproducibility of this survey, a systematic literature review (SLR) was conducted following a structured and well-defined protocol. This methodology was designed to identify, filter, and analyze relevant studies on FPGA floorplanning for Dynamic Partial Reconfiguration (DPR) over the past two decades.

A. Search Strategy and Data Sources

The literature search targeted primary studies published between 2006 and 2026 to capture the evolution of FPGA floorplanning methodologies. The following major scientific databases were systematically queried: IEEE Xplore Digital Library, ACM Digital Library, SpringerLink, ScienceDirect (Elsevier) and arXiv (to include recent developments in AI-driven approaches).

Search queries were constructed using Boolean operators to combine keywords related to FPGA architecture, floorplanning, and optimization techniques. The main query used was:

"FPGA" OR "Field-Programmable Gate Array") AND ("Floorplanning" OR "Placement") AND ("Dynamic Partial Reconfiguration" OR "DPR" OR "Bitstream Relocation") AND ("Optimization" OR "Machine Learning" OR "Reinforcement Learning"

B. Inclusion and Exclusion Criteria

To ensure relevance and quality, studies were selected based on clear inclusion and exclusion criteria. Only peer-reviewed papers addressing FPGA floorplanning or placement in dynamic partial reconfiguration (DPR) contexts, and providing experimental evaluation, were included. Works focusing on ASICs, static FPGA designs, or lacking sufficient technical contribution were excluded. The selection process involved removing duplicates, screening titles and abstracts, and performing full-text analysis, resulting in a refined set of studies for detailed review.

C. Data Extraction and Taxonomy Mapping

For each selected study, key information was extracted, including methodology type, optimization objectives, computational complexity, and evaluation metrics. The studies were then categorized according to the proposed three-generation taxonomy: Heuristic, Optimization-based, and Autonomous approaches.

Each method was analyzed using a unified set of performance criteria, including:

- Timing and performance
- Area and resource utilization
- Communication and routing cost

- Reconfiguration overhead
- Power and energy consumption

This structured analysis enabled consistent comparison across methodologies and supported the identification of research trends and open challenges.

IV. Evolution of FPGA Floorplanning Methodologies: A Taxonomy and State-of-the-Art Review

The evolution of floorplanning methodologies for Dynamic Partial Reconfiguration (DPR) reflects the increasing complexity of modern FPGA architectures and the growing demand for runtime adaptability. Over the past two decades, approaches have progressed from manual and heuristic techniques to formal optimization models and, more recently, to AI-driven autonomous systems.

In this work, we propose a structured taxonomy that categorizes these methodologies into three distinct generations based on their level of algorithmic intelligence, hardware awareness, and adaptability. This classification (Table I) provides a unified framework for understanding the progression of the field and identifying emerging research trends.

While early work focused on the physical legality of bitstream alignment (Era 1), middle-period research introduced formal mathematical models to handle heterogeneous resources (Era 2). Currently, the field is transitioning toward autonomous, AI-augmented systems that prioritize predictive management and elastic resource allocation (Era 3).

This taxonomy highlights a clear shift from static, constraint-based design toward dynamic, intelligent, and adaptive floorplanning strategies.

The following subsections describe each generation in detail, highlighting their key methodologies, innovations, and limitations.

A. Era 1: Heuristic and Rule-Based Foundations

This era is primarily characterized by constraint-driven design and limited automation, focusing on physical feasibility rather than optimality.

1) Hierarchical PBlock methodology: Lysaght et al. [22] introduced structured, constraint-driven partitioning using *PlanAhead*. PRRs were created within fixed physical blocks (PBlocks), allowing early estimation of resource usage, timing analysis, and routing isolation. Static regions were implemented first, followed by PRRs containing Reconfigurable Modules (RMs) independently.

2) Automation of resource alignment: OpenPR [24] and GOAHEAD [23] automated AREA GROUP constraints, bus macro placement, and routing alignment. Clock and frame alignment ensured PRR relocatability and predictable timing, simplifying subsequent integration with partial bitstreams.

3) Slot-based and tile abstractions: ReCoBus-Builder [25] introduced fine-grained resource slots enabling modular relocation, while CoPR [26] adopted tile-based abstractions of CLBs, BRAMs, and DSPs to create precise rectangular PRR layouts, facilitating efficient routing and minimal fragmentation.

4) Low-level logic exploration: RapidSmith 2 [27] exposed BEL-level FPGA structures to CAD tools, enabling placement, packing, and routing optimizations while retaining full compatibility with vendor bitstream generation.

This era focused on legal PRR creation, interface consistency, and minimal human intervention, but largely relied on heuristics, fixed templates, and rule-based strategies without formal guarantees on optimality.

This lack of optimality and inability to handle increasing resource heterogeneity necessitated the shift toward the formal mathematical models characterizing Era 2

B. Era 2: Formal Optimization and Analytical Modeling

This era introduces formal optimization techniques, enabling systematic handling of heterogeneous resources and improving solution quality. As FPGA fabrics became heterogeneous, floorplanning evolved toward mathematical modeling, Mixed Integer Linear Programming (MILP), and nonlinear optimization to achieve predictable and high-quality layouts.

1) Nonlinear and integer programming: PRFloor [28] combined heuristic and NLP optimization to reduce inter-module communication cost while respecting heterogeneous resources. DUPRFloor [29] applied MILP formulations for L-shaped PRRs, ensuring no overlaps and optimized interconnect usage. Deak et al. [30] used Columnar Kernel Tessellation and NLIP to align rectangular PRRs to FPGA frames. FLORA [31] treated FPGA resources as weighted grids to minimize wirelength and routing congestion, considering both logic and memory utilization.

2) Model-driven task mapping: Ben Abdelali et al. [32] developed a two-level L1/L2 scheduling framework. L1 analyzed functional dependencies and reconfiguration transitions between tasks, while L2 scheduled bitstream loading and task execution, optimizing runtime overhead and reducing reconfiguration latency.

3) Metaheuristic and joint optimization: Sadeghi et al. [33] and Ding et al. [34] used Genetic Algorithms and Simulated Annealing (Partitioned Sequence Triple, P-ST) to jointly optimize PRR placement, communication, and heterogeneous resource usage, improving layout quality for complex multi-module workloads.

4) High-level synthesis integration: HiPR [35] linked HLS with PR workflows, enabling designers to define partial functions in C/C++ while automatically generating legal PRR placements. HiPlanner produced optimized floorplans in under a second, minimizing wirelength and ensuring PRR alignment with frames.

5) Heterogeneous resource awareness: The PR FP tooll [36] employs a scarcity-prioritized iterative refinement approach using simulated annealing to handle irregular fabric structures, balancing logic, DSP, and memory usage across multiple PRRs.

Era 2 represents the transition from heuristic approaches to formal, resource-aware optimization, providing precise PRR layouts, efficient inter-module communication, and systematic management of heterogeneous FPGA fabrics, forming the foundation for runtime-adaptive reconfiguration.

TABLE I. TAXONOMY OF FPGA FLOORPLANNING METHODOLOGIES ACROSS GENERATIONS (2006–2026)

Generation	Timeline	Core Methodology	Relocation Support	Primary Focus
Era 1	2006–2015	Heuristic / Rule-based	1D Columnar / Manual	Physical Legality
Era 2	2016–2022	Formal Optimization	Block-Level / MILP	Timing & Congestion
Era 3	2022–2026	AI-Augmented / RL	**2D Shape-Adaptive**	**Autonomous Autonomy**

C. Era 3: Autonomous and AI-Augmented Systems

Era 3 marks a shift toward self-optimizing FPGA systems, where AI techniques are used to address the growing complexity of advanced nodes (e.g., 5nm) and heterogeneous 3D-IC architectures. In contrast to earlier static or design-time approaches, this era integrates intelligence across both runtime adaptation and design-time optimization, structured around three complementary pillars.

1) Reinforcement Learning (RL) for runtime adaptation (Online): Reinforcement Learning is increasingly employed for online decision-making in dynamic environments. In this setting, the FPGA fabric is modeled as an environment in which an agent learns optimal placement or reconfiguration policies through interaction. Such approaches are particularly effective for thermal- and power-aware management, enabling real-time relocation of modules based on sensor feedback. For instance, Yeonjin & Kigarura [37] explore RL for autonomous edge intelligence, while Muralidharan [38] integrates embedded RL logic to trigger DPR events without disrupting ongoing static operations.

2) Evolutionary and metaheuristic search for DSE (offline): Evolutionary and metaheuristic methods have matured into powerful tools for autonomous Design Space Exploration (DSE). Building on earlier Genetic Algorithm approaches, these techniques address the NP-hard nature of floorplanning by exploring large solution spaces and converging toward near-optimal layouts[39]. Unlike RL-based methods, they are typically applied offline during compilation to generate optimized, shape-adaptive floorplans for complex and heterogeneous workloads such as CNN pipelines or signal processing chains.

3) Neural networks and generative AI (Hybrid): Recent advances incorporate neural models, including Graph Neural Networks (GNNs) and Large Language Models (LLMs), to further automate and accelerate the design process. GNNs are used for predictive analysis, such as estimating routing congestion or timing violations prior to finalizing a floorplan. In parallel, LLMs enable generative hardware design flows, translating high-level descriptions into HDL code with embedded placement constraints. For example, frameworks like ResBench [40] assess the quality of LLM-generated designs, while Boudjadar et al. [41] leverage AI-driven DSE for dynamic HW/SW partitioning between Processor Systems (PS) and Programmable Logic (PL) based on runtime requirements.

4) Architectural enablers: Hierarchical and nested PRRs: These algorithmic advances are supported by evolving FPGA architectures that enhance reconfiguration flexibility. In particular, hierarchical and nested Partial Reconfigurable Regions (PRRs) [42] replace traditional flat layouts, enabling finer-grained and independent reconfiguration within subregions (see Fig. 4). This structural elasticity reduces bitstream size, improves scalability, and reinforces the overall autonomy of the system.

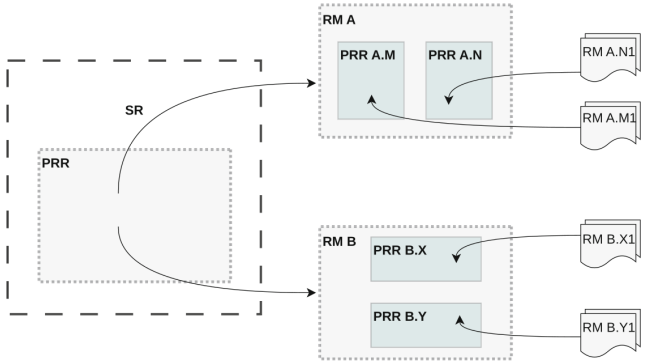


Fig. 4. Hierarchical nested PRR flow exhibiting parent-child reconfiguration layers [42].

The systematic evolution across the three eras is summarized in Table II. A closer examination of the *Major Innovation* and *Primary Metrics* columns highlights a clear progression in the maturity of the field. Era 1 established the practical feasibility of FPGA floorplanning through rule-based and one-dimensional bitstream relocation techniques. Era 2 advanced the state of the art by introducing formal optimization models to address resource heterogeneity and enforce design constraints. Most notably, Era 3 reflects a shift toward spatio-temporal autonomy, where the focus moves from static objectives such as timing closure to dynamic, adaptive metrics driven by runtime considerations, including thermal and power management enabled by AI-based methods.

Overall, the evolution of floorplanning methodologies demonstrates a progressive transition from manual and heuristic approaches to formal optimization and, ultimately, to AI-driven autonomous systems. While each generation addresses specific limitations of its predecessors, emerging challenges such as scalability, heterogeneity, and runtime adaptability continue to drive innovation in this field.

V. CRITICAL ANALYSIS AND RESEARCH GAPS

Despite significant progress across the three generations of FPGA floorplanning methodologies, several fundamental limitations remain unresolved. While heuristic approaches ensure feasibility, optimization-based methods improve solution quality, and AI-driven techniques enhance adaptability, none of these paradigms fully address the combined challenges of scalability, heterogeneity, and runtime dynamism.

This section provides a critical analysis of these limitations and identifies key research gaps that must be addressed to enable next-generation reconfigurable systems.

TABLE II. CHRONOLOGICAL EVOLUTION OF FPGA FLOORPLANNING METHODOLOGIES: FROM HEURISTIC RULES TO AI-AUGMENTED AUTONOMY

Year	Reference	Method / Complexity	Execution Context & Cost	Constraint Guarantees	Major Innovation	Primary Optimization Metrics
ERA 1: HEURISTIC AND RULE-BASED FOUNDATIONS						
2006	Lysaght et al. [22]	Heuristic	High (Offline)	Heuristic constraints	Hierarchical PBlocks (Manual)	Timing Isolation & Resource Est.
2008	ReCoBus-Builder [25]	Heuristic	Low (Offline)	Heuristic constraints	Slot-Based Abstraction 1D	Reconfiguration Overhead
2011	OpenPR [24]	Heuristic	Low (Offline)	Heuristic constraints	Automated AREA GROUP	Timing Closure
2012	GOAHEAD [23]	Heuristic	Low (Offline)	Heuristic constraints	Clock/Frame Alignment 1D	Manual constraint reduction
2015	RapidSmith 2 [27]	Heuristic	Low (Offline)	Heuristic constraints	BEL-level Exploration	Resource Density
ERA 2: FORMAL OPTIMIZATION AND ANALYTICAL MODELING						
2016	PRFloor [28]	NP-hard (NLP)	High (Offline)	Formal guarantees	Hybrid Heuristic/NLP	Communication Cost & Wirelength
2017	Ben Abdelali et al. [32]	Heuristic	Moderate (Offline)	Heuristic constraints	Two-Level L1/L2 Model	Reconfiguration Latency
2019	FLORA [31]	Heuristic	Low (Offline)	No formal guarantees	Weighted Resource Grid	Routing Congestion
2020	DUPRFloor [29]	NP-hard (MILP)	High (Offline)	Formal guarantees	L-shaped PRR Formulation	Overlap Prevention
2022	Ding et al. [34]	NP-hard (SA)	High (Offline)	No formal guarantees	P-ST Joint Optimization	Placement Success Rate
ERA 3: AUTONOMOUS AND AI-AUGMENTED SYSTEMS						
2022	Fahmy et al. [42]	Heuristic	Low (Runtime)	Heuristic constraints	Hierarchical/Nested PRRs	Fine-grained Adaptability
2025	Gu et al. [43]	Heuristic	Low (Runtime)	No formal guarantees	Big.Little Spatio-temporal Slots	Resource Utilization Efficiency
2025	Boudjadar et al. [41]	Heuristic (DSE)	Moderate (Offline)	No formal guarantees	HW/SW Co-Design (CNN)	Throughput & Scaling
2026	Yeonjin & Kigarura [37]	Polynomial	Near Real-time	No formal guarantees	RL-based Thermal Agents	Power & Thermal Management
2026	Muralidharan [38]	Polynomial	Real-time	No formal guarantees	Embedded AI Control Logic	Non-intrusive DPR triggering
2026	Li et al. [44]	Heuristic	Near Real-time	Heuristic constraints	FEditor (2D Shape-Adaptive)	Resource & Shape Flexibility

A. Scalability vs. Optimality Trade-off

A fundamental limitation across existing methodologies is the trade-off between scalability and optimality. Formal optimization techniques, such as MILP and nonlinear programming (Era 2), provide high-quality and constraint-compliant solutions. However, their computational complexity grows exponentially with design size, making them impractical for large-scale FPGA systems.

In contrast, AI-driven approaches (Era 3), particularly reinforcement learning, can explore large design spaces efficiently and produce solutions in near real-time. However, these methods lack guarantees of feasibility and constraint satisfaction, particularly with respect to strict FPGA architectural rules.

This dichotomy highlights the need for hybrid approaches that combine fast exploration capabilities of AI with the rigor of formal verification and constraint enforcement.

B. Fragmentation in Heterogeneous 3D-ICs

The increasing adoption of heterogeneous 2.5D and 3D FPGA architectures introduces new spatial and thermal challenges that are insufficiently addressed in current literature. Most existing approaches assume a 2D fabric and fail to account for Super Logic Region (SLR) boundaries, interposer constraints, and non-uniform communication costs.

Moreover, thermal effects are becoming a critical factor in high-density FPGA systems. Current floorplanning methods largely ignore thermal-aware placement and dynamic heat distribution, which can significantly impact system reliability and performance.

Future research must integrate spatial, thermal, and communication-aware optimization models to address these emerging challenges.

C. Model-Driven Adaptability

While model-driven approaches (Era 2) effectively capture task dependencies and reconfiguration sequences, most AI-driven methods (Era 3) optimize spatial metrics such as area and wirelength in isolation. This results in a disconnect between functional behavior and physical implementation.

In practical systems, reconfigurable modules exhibit temporal dependencies, where certain tasks are frequently co-located or executed sequentially. Ignoring these relationships leads to suboptimal runtime performance and increased reconfiguration overhead.

Future approaches should integrate task-level dependency models into AI-based floorplanning to enable more context-aware and application-driven optimization.

D. Lack of Standardized Benchmarks and Evaluation Frameworks

A major limitation in the field is the absence of standardized benchmarks and evaluation methodologies. Existing works often rely on custom datasets, proprietary designs, or tool-specific configurations, making it difficult to compare approaches objectively.

This lack of reproducibility hinders fair evaluation and slows progress in the field. Future research should focus on developing open benchmarks and unified evaluation metrics to enable consistent comparison across methodologies.

E. Limited Integration with Industrial CAD Flows

Despite significant academic advances, many proposed methodologies are not fully compatible with industrial FPGA design tools such as Vivado or Quartus. This gap limits their practical adoption and real-world impact.

Bridging the gap between research prototypes and industrial toolchains remains a key challenge for the field.

F. Future Research Directions

Based on the identified gaps, several promising research directions emerge:

- Hybrid AI-Optimization Frameworks: Combining reinforcement learning with formal constraint solvers to ensure both scalability and correctness.
- Thermal-Aware and 3D Floorplanning: Incorporating temperature modeling and SLR-aware placement strategies.
- Application-Aware Optimization: Integrating task dependencies and runtime behavior into floorplanning decisions.
- Standardized Benchmarks: Developing open datasets and evaluation frameworks for fair comparison.
- Toolchain Integration: Ensuring compatibility with industrial FPGA design environments.

These directions are expected to drive the next generation of intelligent and autonomous FPGA systems.

In summary, while significant progress has been achieved, FPGA floorplanning for DPR remains an open and evolving research area. Addressing these challenges will require interdisciplinary approaches that combine optimization theory, machine learning, and hardware-aware design methodologies.

VI. Conclusion

This study has presented a structured taxonomy of the evolution of FPGA floorplanning for Dynamic Partial Reconfiguration (DPR), organized into three distinct eras. The analysis highlights a clear progression from constraint-driven heuristic methods (Era 1), to formal optimization models addressing architectural heterogeneity (Era 2), and finally to AI-driven autonomous systems enabling runtime adaptation (Era 3).

Despite these advances, no single approach fully addresses the combined challenges of scalability, heterogeneity, and dynamic runtime behavior. This survey therefore emphasizes the need for more flexible and context-aware floorplanning strategies that bridge the gap between design-time optimization and runtime adaptability.

Beyond this classification, the study identifies three key research directions:

- Scalability–Optimality Trade-off: A fundamental gap exists between the formal rigor of Era 2 approaches and the exploration efficiency of Era 3 methods. Future solutions are expected to adopt hybrid frameworks that combine formal constraint handling with learning-based decision-making.
- Thermal–Spatial Management in 3D Architectures: With the emergence of advanced nodes and multi-die FPGA systems, floorplanning increasingly serves as a mechanism for runtime thermal regulation. This requires tighter integration between placement decisions and thermal-aware reconfiguration strategies.
- Temporal Awareness in Autonomous Systems: Current approaches often decouple spatial optimization from task-level temporal dependencies. Future work should focus on context-aware methods that jointly consider placement decisions and execution ordering constraints.

In conclusion, the future of reconfigurable computing depends on the integration of intelligent decision-making within FPGA design flows. Progress in standardized benchmarks and improved CAD tool interoperability will be essential to enable the next generation of adaptive and autonomous reconfigurable systems.

References

[1] A. Boutros and V. Betz, *Field-Programmable Gate Array Architecture*. Singapore: Springer Nature Singapore, 2022, pp. 1–47.

[2] V. Pangracious, Z. Marrakchi, and H. Mehrez, *Field Programmable Gate Arrays: An Overview*. Cham: Springer International Publishing, 2015, pp. 43–71.

[3] K. Vipin and S. A. Fahmy, "Fpga dynamic and partial reconfiguration: A survey of architectures, methods, and applications," *ACM Comput. Surv.*, vol. 51, no. 4, 2018.

[4] J.-P. Deschamps, G. D. Sutter, and E. Cantó, *Partial Reconfiguration on Xilinx FPGAs*. Dordrecht: Springer Netherlands, 2012, pp. 435–459.

[5] W. Chouchene, "Vers une reconfiguration dynamique partielle parallèle par prise en compte de la régularité des architectures fpga-xilinx," 2017.

[6] A. Sadeghi, M. Zolfy Lighvan, and P. Prinetto, "Automatic and simultaneous floorplanning and placement in field-programmable gate arrays with dynamic partial reconfiguration based on genetic algorithm," *Canadian Journal of Electrical and Computer Engineering*, vol. 43, no. 4, pp. 224–234, 2020.

[7] J. Cong, M. Romesis, and J. R. Shinnerl, "Fast floorplanning by look-ahead enabled recursive bipartitioning," 2005. [Online]. Available: https://doi.org/10.1145/1120725.1120838

[8] A. Badhoutiya, Z. Jaffer, H. M. Hussein, A. Juyal, M. Mittal, and R. Anand, "Field programmable gate array: An extensive review, recent trends, challenges and applications," in *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2024, pp. 1084–1090.

[9] L. Singhal and E. Bozorgzadeh, "Heterogeneous floorplanner for fpga," in *15th Annual IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM 2007)*, 2007, pp. 311–312.

[10] K. Vipin and S. A. Fahmy, "Fpga dynamic and partial reconfiguration: A survey of architectures, methods, and applications," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–39, 2018.

[11] "Supported/Unsupported Features • Vivado Design Suite User Guide: Dynamic Function eXchange (UG909) • Reader • AMD Technical Information Portal." [Online]. Available: https://docs.amd.com/r/en-US/ug909-vivado-partial-reconfiguration/Supported/Unsupported-Features?tocId=Ilw7PlHTVUODn9xBfPjdKA

[12] S. U. Bhandari, S. Subbaraman, S. Pujari, and R. Mahajan, "Internal dynamic partial reconfiguration for real time signal processing on fpga," *Indian Journal of Science and Technology*, vol. 3, no. 4, pp. 365–368, 2010.

[13] P. Manet, D. Maufroid, L. Tosi, G. Gailliard, O. Mulertt, M. Di Ciano, J.-D. Legat, D. Aulagnier, C. Gamrat, R. Liberati *et al.*, "An evaluation of dynamic partial reconfiguration for signal and image processing in professional electronics applications," *EURASIP Journal on Embedded Systems*, vol. 2008, no. 1, p. 367860, 2009.

[14] A. Alkamil and D. G. Perera, "Towards dynamic and partial reconfigurable hardware architectures for cryptographic algorithms on embedded devices," *IEEE Access*, vol. 8, pp. 221 720–221 742, 2020.

[15] F. Galea, S. Carpov, and L. Zaourar, "Multi-start simulated annealing for partially-reconfigurable fpga floorplanning," in *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2018, pp. 1335–1338.

[16] "Chapter 10 - floorplanning," in *Electronic Design Automation*, L.-T. Wang, Y.-W. Chang, and K.-T. T. Cheng, Eds. Morgan Kaufmann, 2009, pp. 575–634.

[17] F. Mao, Y. Ma, N. Xu, S. Liu, Y. Wang, and X. Hong, "Congestion-driven floorplanning based on two-stage optimization," in *2009 IEEE 8th International Conference on ASIC*, 2009, pp. 1298–1301.

[18] A. Safir, B. Haroun, and K. Thulasiraman, "Floorplanning with datapath optimization," in *Proceedings of ISCAS'95 - International Symposium on Circuits and Systems*, 1995, pp. 41–44.

[19] M. Ibro and G. Marinova, "Fpga power consumption optimization methods analysis," in *2023 International Conference on Electromechanical and Energy Systems (SIELMEN)*, 2023, pp. 1–4.

[20] X. Zhang, H. Rabah, and S. Weber, "Dynamic slowdown and partial reconfiguration to optimize energy in fpga based auto-adaptive sopc," in *4th IEEE International Symposium on Electronic Design, Test and Applications (delta 2008)*, 2008, pp. 153–157.

[21] F. Sun, H. Wang, F. Fu, and X. Li, "Survey of fpga low power design," in *2010 International Conference on Intelligent Control and Information Processing*, 2010, pp. 547–550.

[22] P. Lysaght, B. Blodget, J. Mason, J. Young, and B. Bridgford, "Enhanced architectures, design methodologies and cad tools for dynamic reconfiguration of xilinx fpgas," in *2006 International Conference on Field Programmable Logic and Applications*. IEEE, 2006, pp. 1–6.

[23] C. Beckhoff, D. Koch, and J. Torresen, "Go ahead: A partial reconfiguration framework," in *2012 IEEE 20th International Symposium on Field-Programmable Custom Computing Machines*. IEEE, 2012, pp. 37–44.

[24] A. A. Sohanghpurwala, P. Athanas, T. Frangieh, and A. Wood, "Openpr: An open-source partial-reconfiguration toolkit for xilinx fpgas," in *2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum*. IEEE, 2011, pp. 228–235.

[25] D. Koch, C. Beckhoff, and J. Teich, "Recobus-builder—a novel tool and technique to build statically and dynamically reconfigurable systems for fpgas," in *2008 International conference on field programmable logic and applications*. IEEE, 2008, pp. 119–124.

[26] K. Vipin and S. A. Fahmy, "Mapping adaptive hardware systems with partial reconfiguration using copr for zynq," in *2015 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*. IEEE, 2015, pp. 1–8.

[27] T. Haroldsen, B. Nelson, and B. Hutchings, "Rapidsmith 2: A framework for bel-level cad exploration on xilinx fpgas," in *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, 2015, pp. 66–69.

[28] T. D. Nguyen and A. Kumar, "Prfloor: An automatic floorplanner for partially reconfigurable fpga systems," in *FPGA 2016 - Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. Association for Computing Machinery, Inc, 2 2016, pp. 149–158.

[29] J. Wang, Y. Kang, W. Wu, G. Xing, and L. Tu, "Duprfloor: Dynamic modeling and floorplanning for partially reconfigurable fpgas," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 8, pp. 1613–1625, 2020.

[30] N. Deak, O. Cret, and H. Hedesiu, "Efficient fpga floorplanning for partial reconfiguration-based applications," in *2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 2019, pp. 309–309.

[31] B. B. Seyoum, A. Biondi, and G. C. Buttazzo, "FLORA: FLoorplan Optimizer for Reconfigurable Areas in FPGAs," *ACM Transactions on Embedded Computing Systems*, vol. 18, no. 5s, pp. 1–20, Oct. 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3358202

[32] A. B. Abdelali, M. Hannachi, M. N. Krifa, H. Rabah, and A. Mtibaa, "High-level design flow and environment for fpga-based dynamic partial reconfiguration," *International Journal of Electronics*, vol. 104, no. 8, pp. 1254–1284, 2017. [Online]. Available: https://doi.org/10.1080/00207217.2017.1293172

[33] A. Sadeghi, M. Z. Lighvan, and P. Prinetto, "Automatic and simultaneous floorplanning and placement in field-programmable gate arrays with dynamic partial reconfiguration based on genetic algorithm," *Canadian Journal of Electrical and Computer Engineering*, vol. 43, no. 4, pp. 224–234, 2020.

[34] B. Ding, J. Huang, J. Wang, Q. Xu, S. Chen, and Y. Kang, "Task modules Partitioning, Scheduling and Floorplanning for Partially Dynamically Reconfigurable Systems Based on Modern Heterogeneous FPGAs," Dec. 2022, arXiv:2212.05397 [eess]. [Online]. Available: http://arxiv.org/abs/2212.05397

[35] Y. Xiao, A. Hota, D. Park, and A. Dehon, "Hipr: High-level partial reconfiguration for fast incremental fpga compilation," Tech. Rep. [Online]. Available: https://github.com/icgrp/hipr

[36] P. Goswami and D. Bhatia, "Automated floorplanning for partially reconfigurable designs on heterogenrous fpgas," Tech. Rep.

[37] K. Yeonjin and M. Kigarura, "Ai-augmented dynamic partial reconfiguration for adaptive edge intelligence in fpga-based systems," *SCCTS Journal of Embedded Systems Design and Applications*, vol. 3, no. 1, pp. 20–27, 2025.

[38] J. Muralidharan and D. Abdullah, "Ai-augmented runtime reconfiguration for energy-aware fpga-based edge computing systems," *SCCTS Transactions on Reconfigurable Computing*, vol. 3, pp. 48–59, 2026.

[39] G. Montanaro, A. Galimberti, and D. Zoni, "Omega: A hardware-software framework for complete design space exploration of fpga-based heterogeneous multi-core socs," *IEEE Transactions on Computers*, 2025.

[40] C. Guo and T. Zhao, "Resbench: Benchmarking LLM-Generated FPGA Designs with Resource Awareness," *Proceedings of International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies (HEART '25)*, vol. 1, 2025, [Online]. [Online]. Available: http://arxiv.org/abs/2503.08823

[41] J. Boudjadar, S. U. Islam, and R. Buyya, "Dynamic FPGA Reconfiguration for Scalable Embedded Artificial Intelligence (AI): A Co-Design Methodology for Convolutional Neural Networks (CNN) Acceleration," *Future Generation Computer Systems*, vol. 169, p. 8, 2025.

[42] S. A. Fahmy and K. B. Iyer, "Dynamic and Partial Reconfiguration of FPGAs," in *Handbook of Computer Architecture*, A. Chattopadhyay, Ed. Springer Nature Singapore, 2022, pp. 1–24. [Online]. Available: https://doi.org/10.1007/978-981-15-6401-7_51-1

[43] J. Gu, H. Wang, X. Guo, M. Schulz, and M. Gerndt, "Versaslot: Efficient fine-grained fpga sharing with big.little slots and live migration in fpga cluster," in *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, 2025, pp. 1–7.

[44] Y. Li, Y. Chen, Z. Xu, Y. Wang, H. Jiang, and K. Li, "Feditor: Consecutive Task Placement with Adjustable Shapes Using FPGA State Frames," *IEEE Transactions on Parallel and Distributed Systems*, vol. 37, pp. 1–14, 2026.

SentimentPulse: A Concurrent Multi-Platform System for Social Media Sentiment Monitoring with LLM-Based Interpretation

Gabriel A. León-Paredes, Erika C. Villa-Quishpi, Jorge E. Márquez-Chávez, Erick F. Zhigue-Granda
Research Group in Cloud Computing, Smart Cities and High Performance Computing
Universidad Politécnica Salesiana, Cuenca, Ecuador

***Abstract*—Monitoring public perception on social media is increasingly important for detecting reputational risks and communication opportunities in rapidly evolving digital environments. However, operational sentiment monitoring remains challenging due to platform fragmentation, heterogeneous data formats, and the need to generate interpretable reports quickly when specialized analysts are not available. This study presents SentimentPulse, a web-based system for multi-platform sentiment monitoring driven by a user-defined query. The system integrates concurrent data extraction from X, Facebook, LinkedIn, and Instagram with large language model (LLM) based sentiment classification and automated executive storytelling generation. The architecture combines parallel scraping processes for data acquisition with multithreaded LLM inference to improve throughput, while structured persistence enables job tracking and cross-platform analysis. The system operates under practical constraints, including platform-specific access limitations, dynamic content availability, and dependency on external LLM services, which may introduce variability in response times and outputs. The evaluation is conducted under controlled experimental conditions using fixed query limits and asynchronous execution settings, and results should be interpreted within these operational boundaries. Experimental results from two anonymized case studies demonstrate the effectiveness and operational performance of the approach. In the first case study, the system processed 1032 social media interactions and produced a sentiment distribution of 49.0% positive, 36.6% negative, and 14.2% neutral, with a manual validation accuracy of 0.88. In the second case study, the pipeline processed 1121 records, with parallel scraping accounting for the majority of the runtime and LLM inference achieving a throughput of 12.08 items per second. These results show that combining concurrent multi-platform extraction with LLM-based interpretation enables practical and interpretable social listening workflows, while highlighting the importance of considering system-level constraints when deploying such solutions in real-world environments.**

***Keywords*—Sentiment analysis; social media monitoring; large language models; social listening systems; multi-platform analytics**

I. Introduction

Sentiment analysis is a central aspect of opinion mining. It seeks to automatically identify the polarity and subjectivity present in text, with applications in perception monitoring, customer service, market research, and social analysis [1], [2], [3], [4]. In recent years, the availability of user-generated content on social media has made this task a practical tool for understanding collective dynamics in (almost) real-time, driving its adoption in scenarios where rapid response is critical. Classic examples show that aggregated sentiment signals on social platforms can correlate with external phenomena such as financial markets or electoral behavior [5], [6], [7].

However, bringing sentiment analysis into an operational environment presents challenges that cannot be solved with a manual approach or isolated tools. First, data are distributed across multiple platforms, each with different formats, access mechanisms, and interaction dynamics, making it difficult to collect content in a consistent and comparable manner across networks [8], [9]. Second, in corporate communication or marketing contexts, interest is not limited to overall percentages: early detection of reputational risks and shifts in tone is required, especially during sensitive events where conversations can escalate quickly [10], [11]. Consequently, it becomes important not only to classify sentiment but also to explain patterns and differentiate results by platform in order to support concrete decisions. Despite the extensive literature on sentiment classification models, comparatively fewer studies focus on the design and evaluation of complete operational systems that integrate multi-platform data acquisition, scalable processing, and decision-oriented interpretation.

This project proposes a unified web application for multi-platform sentiment monitoring around a user-defined topic. The solution integrates (1) concurrent extraction from X, Facebook, LinkedIn, and Instagram, (2) consolidation of content into a common format, (3) global and per-platform sentiment classification, and (4) automatic generation of narrative summaries that interpret quantitative sentiment indicators to support executive-level decision making. To improve robustness in the face of informal language, regional variation, and changing context, a large-scale language model (LLM) is incorporated to assign sentiment labels and produce short explanations for each text. This is relevant because recent evidence suggests that, although LLMs can be competitive in simple tasks and few-shot scenarios, their performance varies when more complex phenomena or additional structure are required, making prompt design and careful evaluation important [12], [13]. Additionally, recent studies have evaluated conversational models (e.g., ChatGPT) as sentiment classifiers, reporting competitive results in certain settings but also variability depending on the prompt, domain, and task formulation, which motivates validation and stability control practices in real-world deployments [14], [15].

The main contributions of this work are: 1) a concurrent hybrid pipeline that combines multiprocessing for extraction

and multi-threading for LLM inference, 2) a persistence and traceability architecture based on MongoDB that enables status and history tracking by topic, and 3) a web interface that integrates search, visualization, data exploration, and executive narrative within a single workflow.

The remainder of this study is organized as follows: Section II reviews related work; Section III presents the methodology and architecture; Section IV reports performance and quality results; and Section V concludes with recommendations.

II. Related Work

Social media sentiment analysis has evolved from lexicon-based and supervised machine learning methods to deep learning and, more recently, large language model (LLM) based approaches. Nip and Berthelier (2024) [16] summarize this evolution and emphasize that sentiment analysis in social platforms extends beyond polarity detection by incorporating multi-modality, temporal dynamics, interaction structures, network effects, and sentiment propagation. They also outline operational challenges that remain prominent in practice, including high-volume data streams, linguistic heterogeneity, fragmented context, and noisy or automated content. Importantly, they discuss limitations that are particularly relevant when LLMs are used in deployed systems, such as output variability, interpretability constraints, bias due to alignment, and reproducibility concerns.

A substantial line of work focuses on improving sentiment classification quality with neural architectures. Katalinić and Dunđer (2025) [17] study sentiment analysis and anomaly detection in crisis-related tweets using 189,626 English-language messages from the 2023 Turkey–Syria earthquake. Their BERT-based sentiment classifier reports 91% precision, 85% recall, and 88% F1, outperforming a logistic regression baseline (54% F1). For anomaly detection, their autoencoder achieves 89% accuracy and 88% precision, exceeding an isolation forest baseline (69% precision, 64% recall), and ablations indicate that attention yields statistically significant gains ($p = 0.011$). While these results demonstrate strong neural performance for crisis monitoring, the scope is limited to a single platform and does not address multi-platform acquisition or end-to-end operational integration.

Complementing individual studies, Sharma et al. (2025) [18] review 178 papers published between 2014 and 2024 and categorize sentiment tasks into document-level, sentence-level, aspect-based, multi-lingual, multi-modal, and emotion-related settings. Their synthesis reports that transformer-based models, particularly BERT variants, typically reach 88–96% accuracy and often exceed 85% F1 depending on dataset and task difficulty, and they note that LLM-based approaches can surpass 96% accuracy in some domains. Despite these advances, the review highlights persistent challenges, such as sarcasm/irony, domain adaptation, multilingual variation, dataset bias, computational cost, and limited transparency, and observes that fewer works study system-level concerns such as real-time pipelines and deployment constraints.

In addition to model-centric approaches, recent work has explored the use of web-based systems for sentiment analysis in social media environments. Parra-Zambrano and León-Paredes (2024) [14] present a system that integrates natural language processing techniques and large language models, including ChatGPT, to analyse public opinion on political figures in Ecuador. Their approach achieves accuracy levels of up to 95% in classifying sentiments into positive, neutral, and negative categories, demonstrating the effectiveness of combining NLP pipelines with LLM-based components in real-world social media scenarios. These results highlight the potential of hybrid architectures that incorporate LLMs within broader analytical pipelines, while also emphasizing the importance of system-level design for handling heterogeneous and large-scale social media data.

Stability and reproducibility are increasingly recognized as deployment-critical properties. Ouyang et al. (2024) [15] analyze ChatGPT-based sentiment analysis through an AI quality management lens, distinguishing operational uncertainty from model robustness. Using three-way sentiment labels on 983 Amazon review samples and 1101 SST samples, they show that timing and model-version changes can lead to different confusion matrices on the same test set across API releases (e.g., 2023-06, 2023-12, 2024-01), complicating reproducibility. They also evaluate adversarial perturbations (typo, synonym, homoglyph, homophone) with accuracy and attack success rate (ASR), reporting that synonym substitution is the strongest perturbation and that shorter texts tend to be more vulnerable than longer reviews. These results motivate stability-aware practices such as prompt standardization, version tracking, and continuous monitoring in LLM-based pipelines.

Overall, prior work either (1) surveys the methodological landscape and challenges [16], [18], (2) improves model-level performance primarily within a single platform or benchmark setting [17], or (3) evaluates LLMs as sentiment analysers with an emphasis on prompting, robustness, and stability [14], [15]. In contrast, our work focuses on an end-to-end, deployable social listening system that integrates concurrent multi-platform acquisition (X, Facebook, LinkedIn, and Instagram), LLM-based sentiment labelling with per-item explanations, persistence for traceability (corpus, job status, and aggregated analytics), and an executive-oriented storytelling layer for cross-platform interpretation. This system-level contribution targets operational constraints—throughput, observability, and comparability across networks—that are only partially addressed in model-centric studies.

III. Methodology

This section describes the implementation of SentimentPulse at a system level to support understanding and reproducibility. The architecture is presented through its layered design, the processing pipeline, the concurrency strategy, and the REST workflow that allows the frontend to trigger jobs, monitor progress, and retrieve results.

A. Layered System Design

SentimentPulse is organized as a layered architecture connected through a REST API. Fig. 1 summarizes the main components and their interactions, including the API endpoints, background execution model, parallel data extraction

modules, large language model interpretation, natural language processing and analysis, and the database persistence layer.

At the presentation layer, a Vue 3 single-page application provides four main views: search, progress tracking, dashboard visualization, and data exploration. Through these interfaces, users can initiate sentiment analysis jobs for a given topic and consume JSON responses generated by the backend to render metrics, charts, and narrative summaries of the results.

The application logic is exposed through a FastAPI backend that implements the REST interface and manages asynchronous execution of the analysis pipeline. Jobs are executed using background tasks, allowing the system to process long-running operations without blocking the API. During execution, job progress and status information are stored in the persistence layer, enabling polling-based monitoring and improving system observability. Table I summarizes the REST endpoints used in the system workflow.

TABLE I. REST ENDPOINTS USED BY SENTIMENTPULSE

Endpoint	Method	Purpose
`/scrape`	POST	Start a job in background execution
`/status/topic`	GET	Retrieve per-platform progress and phase status
`/results/topic`	GET	Retrieve aggregated outputs and storytelling
`/history`	GET	Retrieve previous runs from stored analyses
`/cancel`	POST	Mark a job as cancelled in the database
`/history/topic`	DELETE	Delete analysis and associated posts for a topic

At the orchestration level, the central function `run_pipeline` coordinates the complete workflow of the system. This component manages the transition between processing phases, aggregates outputs obtained from different social media platforms, and transforms them into a unified tabular schema that allows consistent downstream analysis.

The data extraction stage collects posts and comments related to a user-defined topic. To achieve this, the system launches one Playwright-based scraper per social media platform. Each scraper operates with persistent sessions and collects relevant textual content while preserving metadata needed for subsequent analysis.

Once the textual data has been collected, the interpretation stage applies a large language model to each text unit. The model produces both a sentiment label and a short explanation, which improves interpretability and allows the system to provide traceable reasoning for each classification.

Following interpretation, the NLP and analysis stage processes the labeled texts using a traditional natural language processing pipeline. This stage extracts salient terms and generates aggregated indicators, including global sentiment metrics and platform-specific distributions. These results are then used to produce narrative summaries that facilitate high-level interpretation of the analysis outcomes.

Finally, all intermediate and final outputs are stored in MongoDB. The database maintains the structured corpus, job status information, execution times, and aggregated analytical results. This persistence layer enables efficient retrieval of past analyses and supports the reproducibility and traceability of the sentiment monitoring process.

It is important to note that the proposed system should be interpreted as a research-oriented prototype designed to evaluate system-level integration strategies rather than a fully optimized production platform. While it supports real-world data acquisition and analysis workflows, aspects such as large-scale deployment and long-term operational robustness are beyond the scope of the current implementation.

B. Pipeline Phases

The processing pipeline is executed in sequential phases in order to separate extraction, interpretation, and aggregation tasks. These phases are reflected in job status updates to support real-time monitoring of the system. The execution state is tracked in MongoDB through the `job_status` collection, which stores per-platform progress information such as *current*, *total*, and *status*. In addition, the progress of the LLM interpretation stage is tracked through the field `llm_status`.

The first phase corresponds to the extraction stage. The orchestrator launches a multiprocessing pool with one process per platform, where each process runs a Playwright scraper. These scrapers collect posts and comments related to the target topic and return normalized records containing the platform identifier, the post content, and the associated comment content.

Once the content from all platforms has been collected, the pipeline enters the interpretation and sentiment labelling phase. After consolidation into a DataFrame, the system performs concurrent API calls to the large language model using a thread pool in order to maximize throughput during the I/O-bound inference process. Each record is processed by the model and mapped to a fixed set of sentiment labels: *Positive*, *Neutral*, or *Negative*. The model also generates a short natural-language justification, which is stored in the field `explanation_llm`. Outputs that do not conform to the expected format or fail during processing are mapped to the label *Error*.

For each record, the LLM receives an instruction to assign a discrete polarity label and generate a short justification grounded only in the textual evidence present in the message. The explanation, therefore, reflects linguistic signals such as supportive language, expressions of gratitude, celebration, criticism, insults, frustration, or neutral informational tone, while avoiding assumptions beyond the text itself. In practice, the model assigns the label *Positive* when the message conveys approval, encouragement, pride, or favorable outcomes; *Negative* when it expresses disapproval, anger, accusations, mockery, or hostile language; and *Neutral* when the content is primarily descriptive, interrogative, informational, or lacks clear affective cues. These explanations are later aggregated during the storytelling stage to summarize dominant patterns and cross-platform differences.

After sentiment interpretation, the pipeline proceeds to the classic NLP and term extraction phase. During this stage, the system applies text cleaning and tokenization and computes salient terms using frequency-based signals. These terms support quick corpus inspection and reporting and are exported together with the sentiment outputs in the required CSV structure.

The following phase corresponds to aggregation and storytelling. In this stage, the system computes global and per-

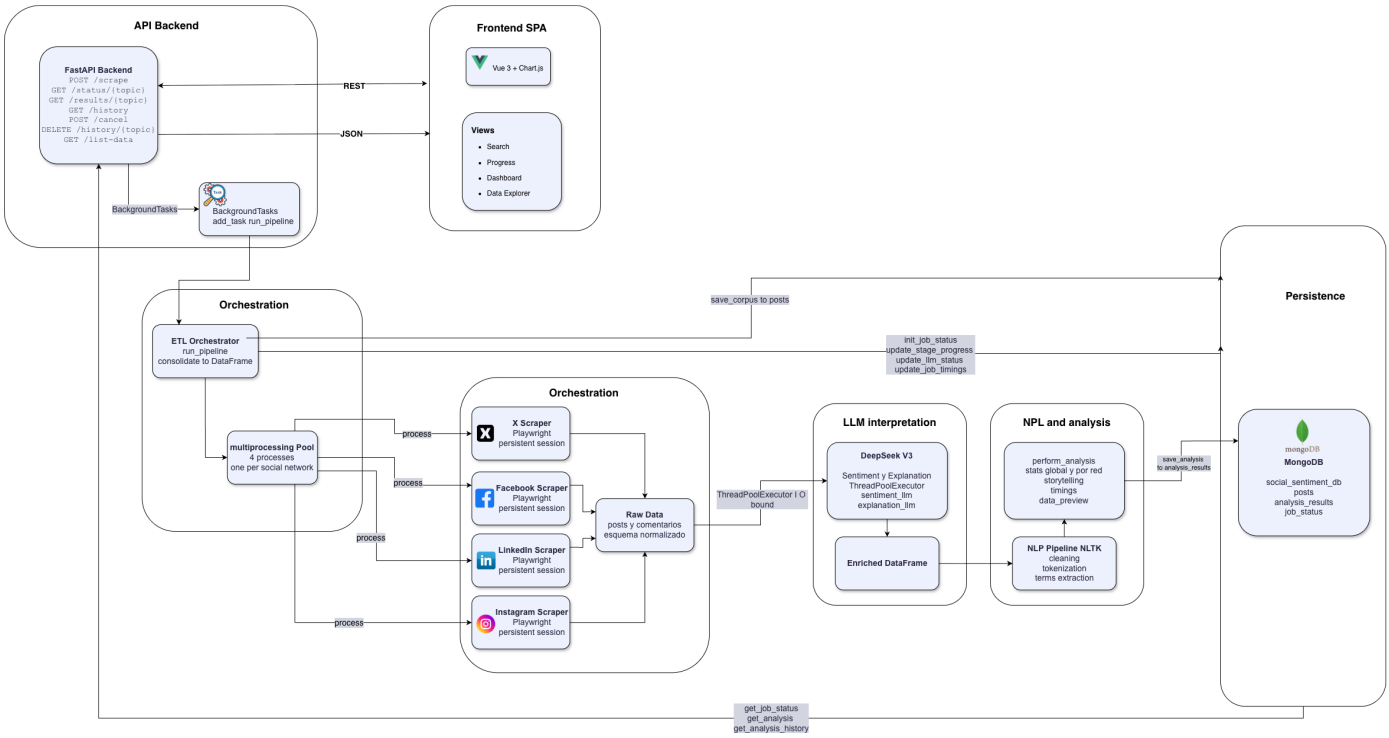


Fig. 1. Layered architecture of SentimentPulse with REST workflow, background execution, parallel scraping, LLM interpretation, NLP and analysis, and database persistence.

platform sentiment counts and percentages, selects representative examples from the dataset, and generates a narrative report intended for executive interpretation. This storytelling component summarizes dominant sentiment patterns and highlights differences across social networks.

Finally, the persistence and export phase stores the results of the analysis. The collected corpus is stored in the `posts` collection, job tracking information is updated in `job_status`, and aggregated outputs are stored in `analysis_results`. Additionally, the system exports a CSV deliverable with the required columns: *Platform*, *Post*, *Comment*, *Sentiment*, *Explication*, and *Terms*.

To mitigate the inherent variability of LLM outputs, the system incorporates several design decisions at the interpretation stage. First, model responses are constrained to a fixed label set (*Positive*, *Neutral*, *Negative*) through prompt standardization and post-processing validation. Second, non-conforming or ambiguous outputs are explicitly mapped to an *Error* category to prevent uncontrolled propagation of inconsistent predictions. Third, the system records the status of the LLM inference phase (`llm_status`) to enable traceability and monitoring of potential inconsistencies during execution.

These mechanisms allow the pipeline to maintain consistent downstream processing despite variability in LLM responses, supporting more stable and auditable sentiment analysis results.

C. Concurrency Strategy and Operational Considerations

The pipeline combines different concurrency mechanisms depending on the characteristics of each stage. This hybrid strategy improves overall system throughput while maintaining isolation between platform-specific operations.

During the extraction phase, scraping tasks are executed using multiprocessing. A separate process is created for each platform worker, allowing the scrapers to run independently and preventing failures or delays in one platform from affecting the others. This approach also improves end-to-end latency because multiple platforms can be processed simultaneously.

In contrast, the interpretation phase relies on concurrent LLM inference implemented with a thread pool. Since the interaction with the language model involves network-bound API calls, a multithreading approach increases throughput by allowing multiple requests to be issued while waiting for responses. This design maximizes resource utilization during the I/O-bound inference stage.

From an operational perspective, social media platforms often rely on dynamic content rendering mechanisms such as infinite scrolling, asynchronous loading, and frequent user interface changes. To maintain robustness under these conditions, the scrapers incorporate controlled scrolling strategies, retry mechanisms, and persistent sessions. Failures or interruptions during extraction are tracked through explicit status fields stored in the system database, allowing post hoc auditing and inspection of scraping runs. Finally, from a data protection perspective, the system is designed to process only publicly available textual content from social media posts and comments. No personally identifiable information (PII) is intentionally extracted, stored, or analyzed by the pipeline. The collected data are limited to the textual content required for sentiment analysis and associated platform metadata necessary for processing.

D. Scalability Considerations

The scalability of the proposed system is primarily determined by the concurrency model adopted for data extraction and LLM-based inference. Scraping tasks are parallelized using multiprocessing, with one independent process per social media platform. As a result, the system scales linearly with the number of platforms, provided that sufficient computational resources (CPU cores and memory) are available.

However, scalability is constrained by several factors. First, the scraping stage is dependent on dynamic web content and platform-specific access mechanisms, which introduce variability in response times and limit the degree of parallelism that can be effectively achieved. Second, LLM-based sentiment classification relies on external API calls, making throughput sensitive to network latency and service rate limits. Third, the overall system performance is bounded by the slowest stage in the pipeline, which in practice corresponds to the data extraction phase.

Empirical results presented in Section IV confirm that scraping constitutes the primary bottleneck, accounting for the majority of the total execution time. These observations indicate that, while the system benefits from concurrent execution, its scalability is inherently constrained by external dependencies and I/O-bound operations rather than computational limitations alone.

E. Algorithmic Summary

Algorithm 1 describes the end-to-end workflow executed after the user submits a topic. The input is a user-defined `topic` and a per-platform `limit`. The pipeline outputs (1) a structured corpus stored in MongoDB (`posts`), (2) an aggregated analysis document with metrics and storytelling (`analysis_results`), and (3) a CSV export with the required columns for the deliverable.

Step 2 to Step 4 perform parallel data acquisition across platforms and consolidate the collected posts and comments into a unified schema. Step 5 to Step 6 execute concurrent LLM inference to assign sentiment labels and explanations. Step 7 to Step 8 extract salient terms and compute global and per-platform statistics, which are then used to generate executive storytelling. Finally, the pipeline persists the corpus and analytics to MongoDB and exports the CSV deliverable.

Algorithm 1 SentimentPulse Multi Platform Sentiment Analysis Pipeline

```
procedure SentimentPulsePipeline(topic, limit)
    init_job_status(topic)
    Phase 1: Parallel scraping
    records ← ∅
    for all platform ∈ {X, Facebook, LinkedIn, Instagram} do in parallel
        platformRecords ← scrape_platform(topic, limit)
        records ← records ∪ platformRecords
    end for
    update_llm_status(topic, running)
    Phase 2: LLM interpretation and sentiment labelling
    results ← ∅
    for all record ∈ records do using thread pool
        label, explanation ← llm_classify(record.text)
        results ← results ∪ (record, label, explanation)
    end for
    update_llm_status(topic, completed)
    Phase 3: NLP term extraction
    terms ← extract_salient_terms(results)
    Phase 4: Aggregation and storytelling
    metrics ← compute_sentiment_metrics(results)
    storytelling ← generate_executive_storytelling(metrics, results)
    Phase 5: Persistence and export
    save_corpus(topic, results)
    save_analysis(topic, metrics, storytelling)
    export_csv(topic, results, terms)
end procedure
```

IV. Results

A. Experimental Setup

The experimental evaluation is based on two anonymized execution runs generated by the SentimentPulse system. The resulting data are stored in CSV files that capture both the sentiment-labeled corpus and the execution-time measurements produced during the analysis pipeline. Specifically, the evaluation considers (1) a sentiment-labeled dataset corresponding to *Case Study A*, and (2) an execution-time log corresponding to *Case Study B*.

Case Study A corresponds to a high-visibility sports topic involving a male professional football player associated with the Ecuadorian national team. Such topics typically generate emotionally charged reactions across social media platforms, including expressions of support, celebration, criticism, and rivalry among users.

Case Study B corresponds to a high-salience political topic involving a prominent political figure in the United States. Political discussions of this type frequently produce polarized discourse and rapid shifts in tone across different platforms.

To comply with ethical and confidentiality considerations, all datasets were anonymised before analysis. Consequently, specific person names, entities, or identifying topic strings are not disclosed in this manuscript.

The collected social media content may include multiple languages as well as informal code-switching. The LLM processes the raw text without language filtering and normalizes all outputs into a fixed sentiment label set consisting of *Positive*, *Neutral*, and *Negative*. Any responses that do not conform to the expected format are mapped to the label *Error*.

B. Case Study A: Dataset Statistics and Sentiment Distribution

Table II summarizes the dataset collected for *Case Study A*. In total, 1,032 records (comments/interactions) were processed across four social media platforms—X, Facebook, LinkedIn, and Instagram—corresponding to 107 unique posts. The largest share of records originated from X (433) and Instagram (321), followed by Facebook (244) and LinkedIn (35). One record contained missing platform metadata and is reported as *Unknown* for completeness.

TABLE II. DATASET SIZE BY PLATFORM FOR *CASE STUDY A*

Platform	Records	Unique posts
X	433	49
Instagram	321	32
Facebook	244	10
LinkedIn	35	15
Unknown	1	1
Total	1032	107

The overall sentiment distribution is reported in Table III. Positive sentiment represents the largest share of the dataset (49.0%), followed by negative sentiment (36.6%) and neutral sentiment (14.2%). A single instance (0.1%) was labelled as *Error*, indicating an unclassified output. Such cases may occur due to transient failures or unexpected responses during automated processing, but their negligible frequency suggests that the pipeline maintained stable inference behaviour.

TABLE III. OVERALL SENTIMENT DISTRIBUTION FOR *CASE STUDY A*

Sentiment	Count	Percentage
Positive	506	49.0%
Negative	378	36.6%
Neutral	147	14.2%
Error	1	0.1%
Total	1032	100%

Table IV provides a platform-level breakdown of sentiment labels. The distribution reveals clear cross-platform differences. X concentrates the largest share of negative reactions (221 instances), reflecting the platform's tendency to host more confrontational or critical discourse in this topic. In contrast, Instagram exhibits a predominantly positive distribution (197 positive versus 89 negative), which is consistent with the more supportive or celebratory tone often observed in visual-centric social media interactions. Facebook presents a more balanced distribution, while LinkedIn contains a relatively small number of records, mostly positive or neutral.

TABLE IV. SENTIMENT BY PLATFORM FOR *CASE STUDY A*

Platform	Positive	Neutral	Negative	Error
X	148	61	221	1
Instagram	197	35	89	0
Facebook	132	44	68	0
LinkedIn	28	7	0	0
Unknown	1	0	0	0

Beyond the numerical distribution, the explanations generated by the LLM provide insight into the linguistic drivers behind the assigned polarity labels. Positive classifications were typically justified by supportive or celebratory expressions such as praise, encouragement, or gratitude. Negative classifications frequently corresponded to critical or hostile language, including insults, accusations, or expressions of frustration. Neutral classifications were generally associated with informational statements, factual descriptions, or questions where explicit affective cues were absent.

These qualitative explanations complement the quantitative analysis by clarifying *why* certain platforms concentrate higher levels of negativity or positivity. In particular, they show that the model's decisions are grounded in explicit textual cues rather than opaque or arbitrary outputs, supporting the interpretability of the system's sentiment classification process.

C. Case Study B: Runtime and Performance

Runtime performance was evaluated using the execution logs generated for *Case Study B* (timestamp: 2026-02-05 23:32:21, limit: 10 posts per platform). In this run, the system processed a total of 1121 records (comments/interactions) collected across the monitored platforms.

Table V summarizes the stage-level execution times recorded in the CSV log. The parallel scraping stage required 678.33 s, while the LLM-based sentiment classification stage required 92.77 s. Considering both stages together, the measured processing time for data acquisition and inference was 771.10 s.

TABLE V. PIPELINE PERFORMANCE FOR *CASE STUDY B* (LIMIT: 10 POSTS PER PLATFORM).

Stage	Time (s)	Proportion
Parallel scraping (4 processes)	678.33	88.0%
LLM classification (ThreadPool)	92.77	12.0%
Measured total (Scraping + LLM)	771.10	100%

Based on these measurements, the LLM inference stage achieved an approximate throughput of 12.08 processed items per second (1121 records divided by 92.77 seconds). The results indicate that the dominant contribution to overall runtime originates from the scraping stage, which represents approximately 88.0% of the measured execution time. In contrast, the LLM inference stage accounts for 12.0% of the total runtime.

This distribution reflects the operational characteristics of social media data acquisition, where dynamic web interfaces, pagination, and interaction-driven loading introduce higher latency compared to API-based model inference. Nevertheless, the concurrent execution strategy—multiprocessing for scraping and multi-threading for LLM inference—allows the system to process more than one thousand social interactions within approximately thirteen minutes in this configuration.

Although the reported execution times correspond to a single experimental run, variability in system performance is expected due to factors such as network latency, platform response times, and external API dependencies. In particular, the scraping stage is subject to dynamic content loading and platform-specific constraints, which may introduce fluctuations in execution time across runs.

D. Storytelling Outputs by Case Study

In addition to quantitative metrics, SentimentPulse generates an executive-oriented *storytelling* report for each run. This report aggregates platform-level sentiment distributions and uses the per-item `explanation_llm` field to interpret *why* polarity patterns emerge. The goal is to transform raw sentiment counts into an interpretable narrative that highlights dominant themes and cross-platform contrasts.

a) Case Study A (sports topic): For *Case Study A*, the storytelling report describes the conversation as predominantly positive, consistent with the distribution presented in Section IV-B. Supportive and celebratory messages constitute the main drivers of positive sentiment, whereas negative sentiment is largely associated with criticism and expressions of frustration. The narrative also emphasizes that polarity in this domain is frequently conveyed through short, emotionally charged statements. Differences across platforms are visible in the tone of engagement: some networks concentrate stronger criticism, while others amplify supportive reactions. These observations are grounded in the LLM explanations, which typically reference praise, encouragement, or gratitude for positive labels and derogatory or accusatory language for negative labels.

b) Case Study B (political topic): For *Case Study B*, the storytelling report characterizes the discussion as highly polarized, with frequent alternation between supportive and hostile language. Compared with the sports topic, the discourse exhibits stronger evaluative framing, including allegations, confrontational phrasing, and explicit ideological positioning. Platform-level patterns are also more pronounced in this case, with certain networks displaying more adversarial exchanges while others contain comparatively neutral or informational contributions. The LLM-generated explanations support these observations by identifying cues such as mockery, accusations, or hostile framing for negative sentiment, and endorsement or approval for positive sentiment.

Overall, the storytelling layer complements the quantitative analysis by linking aggregate sentiment proportions to recurring linguistic cues extracted from the `explanation_llm` field. This additional interpretive layer facilitates faster understanding of the results and provides an auditable explanation of cross-platform sentiment dynamics.

E. Human Validation of Sentiment Classification

To obtain a more reliable estimate of sentiment classification quality, a manual validation was conducted on a random sample of 100 comments drawn from the processed dataset. Each comment was independently labelled by a human evaluator using the same sentiment categories employed by the system (*Positive*, *Neutral*, *Negative*). The predicted labels produced by the LLM were then compared against the human annotations.

The model matched the human label in 88% of the evaluated cases (Accuracy = 0.88), indicating a strong level of agreement under the sampled conditions. Table VI reports per-class precision, recall, and F1-score values. The evaluated sample contains instances from all three sentiment classes, allowing for a more comprehensive assessment of model performance across categories.

The confusion matrix in Table VII shows that most disagreements correspond to polarity shifts between *Positive* and *Negative*, which is consistent with known challenges in sentiment analysis when dealing with ambiguous or mixed expressions. A smaller number of errors involve confusion between *Neutral* and the other classes, suggesting that borderline cases with weak affective signals remain difficult to classify consistently.

TABLE VI. MANUAL VALIDATION METRICS (SAMPLE: 100 COMMENTS)

Class	Precision	Recall	F1-score	Support
Positive	0.97	0.92	0.94	36
Negative	0.79	0.87	0.83	30
Neutral	0.88	0.85	0.87	34
Macro avg	0.88	0.88	0.88	100
Weighted avg	0.88	0.88	0.88	100

TABLE VII. CONFUSION MATRIX (GROUND TRUTH VS. PREDICTED) FOR MANUAL VALIDATION (N=100).

	Predicted (LLM)		
Ground truth (human)	Negative	Neutral	Positive
Negative	26	3	1
Neutral	5	29	0
Positive	2	1	33

Beyond the quantitative agreement, the presence of the `explanation_llm` field enables post-hoc inspection of mismatches. When the predicted label differs from the human label, the associated justification can be examined to determine whether the discrepancy arises from linguistic ambiguity (e.g., sarcasm, mixed sentiment, or implicit tone) or from model misinterpretation. This capability enhances interpretability and provides a practical mechanism for refining prompt design and evaluation strategies.

It is important to note that this validation is based on a single annotated sample and one human evaluator. Therefore, the reported metrics should be interpreted as an indicative assessment of model behaviour rather than a definitive generalization of performance across all possible inputs. Expanding the evaluation to larger and multi-annotator datasets constitutes an important direction for future work.

F. Baseline Comparison

To contextualize the performance of the proposed approach, we consider commonly used sentiment analysis baselines such as VADER and TextBlob, which are widely adopted in social media analysis tasks.

These methods have been widely used in sentiment analysis tasks and have shown competitive performance in detecting explicit sentiment polarity, particularly in short and informal texts [19], [20]. However, prior studies have reported limitations when handling contextual nuances, implicit tone, sarcasm, and mixed sentiment expressions, which are common in social media data [21], [22].

In contrast, the LLM-based approach used in SentimentPulse is designed to capture richer contextual information and provide explanatory outputs, which can improve interpretability and robustness in complex linguistic scenarios. While a direct experimental comparison is left for future work, the results obtained in the manual validation suggest that the proposed approach achieves a high level of agreement with human annotations.

V. CONCLUSION

This work presented SentimentPulse, a unified web application that integrates concurrent multi-platform data extrac-

tion, LLM-based sentiment classification with per-item explanations, and executive-oriented storytelling for operational social listening. The proposed architecture combines parallel scraping, concurrent LLM inference, and structured persistence to support end-to-end analysis of social media conversations across multiple platforms within a single workflow.

Across the reported anonymized case studies, the results show that the system can process large volumes of social media interactions while producing interpretable outputs suitable for cross-platform comparison. In the sports-related topic analyzed in Case Study A, the pipeline processed over one thousand comments across four platforms and revealed a predominantly positive conversation, with substantial variations in sentiment distribution between networks. These differences highlight the importance of analyzing sentiment both globally and at the platform level. The inclusion of the `explanation_llm` field further improves auditability by enabling sentiment labels to be inspected together with short text-grounded justifications.

The runtime analysis in Case Study B shows that the end-to-end pipeline can process more than one thousand interactions in approximately thirteen minutes under the evaluated configuration. The measurements indicate that data extraction remains the dominant cost in the pipeline, accounting for the majority of the execution time, whereas LLM-based inference represents a smaller share of the runtime. A small-scale manual validation achieved an accuracy of 0.85, with strong F1-scores for positive and negative classes, suggesting that the model can reliably detect polarity when sentiment cues are explicit.

Despite these results, several limitations remain. Scraping reliability depends on platform interface changes and access constraints, which may affect both the completeness and latency of data acquisition. Additionally, LLM outputs may vary with prompt formulation and domain shifts, making continuous monitoring and evaluation necessary for operational deployments.

Beyond these limitations, this work highlights the importance of system-level integration in modern sentiment analysis applications. While recent advances in LLMs have significantly improved classification capabilities, their effective use in real-world scenarios depends on how they are embedded within scalable, observable, and interpretable pipelines. In this context, SentimentPulse demonstrates that combining concurrent data acquisition, structured processing, and LLM-based interpretation can support practical social listening tasks under realistic constraints.

Future work should focus on improving extraction efficiency through techniques such as incremental collection and caching, as well as expanding the human validation protocol with larger and more balanced samples to obtain more stable performance estimates across all sentiment classes. These improvements would further strengthen the robustness and scalability of the proposed system for real-world social listening scenarios.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, pp. 1–135, 01 2008.

[2] B. Liu, *Sentiment Analysis and Opinion Mining*, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[3] D. A. Andrade-Segarra, G. A. Le *et al.*, "Deep learning-based natural language processing methods comparison for presumptive detection of cyberbullying in social networks," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021.

[4] G. A. León-Paredes, L. A. Alba-Narváez, S. M. Torres-Cordero, and C. M. Buestan-Villa, "A multilingual sentiment analysis system for tiktok comments in spanish using roberta and lstm," in *International Conference on Information Technology & Systems*. Springer, 2025, pp. 210–219.

[5] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, Mar. 2011.

[6] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 2010, pp. 178–185, conference held in Washington, DC, United States (May 23–26, 2010).

[7] G. A. León-Paredes, B. L. Padilla-Viñanzaca, S. F. Guamán-Torres, and K. M. Parraga-Riera, "A real-time political sentiment analysis system using deepseek-r1 on multiplatform social media data," in *International Conference on Communication and Applied Technologies*. Springer, 2025, pp. 33–42.

[8] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," *AI & SOCIETY*, vol. 30, pp. 89–116, 2015, published online: 26 July 2014.

[9] A. Diwali, K. Saeedi, K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain, "Sentiment analysis meets explainable artificial intelligence: A survey on explainable sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 837–846, 2023.

[10] L. Ott and P. Theunissen, "Reputations at risk: Engagement during social media crises," *Public Relations Review*, vol. 41, no. 1, pp. 97–102, Mar. 2015.

[11] R. T. Rust, W. Rand, M.-H. Huang, A. T. Stephen, and T. Chabuk, "Real-time brand reputation tracking using social media," *Journal of Marketing*, vol. 85, no. 4, pp. 21–43, 2021.

[12] W. Zhang, Y. Deng, B. Liu, S. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 3881–3906. [Online]. Available: https://aclanthology.org/2024.findings-naacl.246/

[13] X. Gu, X. Chen, P. Lu, and Z. Li, "Agcvt-prompt: An automatic general chain-of-thought verbalizer template prompt learning method for sentiment classification," *Engineering Applications of Artificial Intelligence*, vol. 135, p. 107907, 2024.

[14] B. E. Parra-Zambrano and G. A. León-Paredes, "A Web Approach for the Extraction, Analysis, and Visualization of Sentiments in Social Networks Regarding the Public Opinion on Politicians in Ecuador Using Natural Language Processing and High-Performance Computing Tools," in *Information Technology and Systems*, Á. Rocha, C. Ferrás, J. Hochstetter Diez, and M. Diéguez Rebolledo, Eds. Cham: Springer Nature Switzerland, 2024, pp. 457–466.

[15] T. Ouyang *et al.*, "Stability analysis of chatgpt-based sentiment analysis," *Electronics*, vol. 13, no. 24, p. 5043, 2024.

[16] J. Y. M. Nip and B. Berthelier, "Social media sentiment analysis," *Encyclopedia*, vol. 4, no. 4, pp. 1590–1598, 2024.

[17] J. Katalinić and I. Dunđer, "Neural network-based sentiment analysis and anomaly detection in crisis-related tweets," *Electronics*, vol. 14, no. 11, p. 2273, 2025.

[18] N. A. Sharma *et al.*, "A systematic review of sentiment analysis: Tasks, applications, and deep learning techniques," *International Journal of Data Science and Analytics*, vol. 19, no. 3, pp. 351–388, 2025.

[19] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2014.

[20] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[21] E. Cambria, S. Poria, R. Bajpai, and B. Schuller, "Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings," *Proceedings of AAAI*, 2017.

[22] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, pp. 5731–5780, 2022.

Improving Performance and Accuracy in Decision Trees: A Literature Survey on Impurity Functions

Abed Alsulami, Prof. Reda Khalifa, Dr. Wajdi Alghamdi
Information Technology Department, King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Decision tree algorithms have remained some of the most popular tools for supervised learning tasks. This has been because of their comprehensibility, malleability, and robustness to accommodate varying types of data. Nevertheless, their accuracy and performance are dependent on the use of impurity measures to guide the process of tree splits. The standard impurity measures, such as the Gini impurity, Entropy, and Classification Error, tend to face problems as the complexity, imbalance, and noise levels are raised. This often gives rise to overfitting. An examination of the disadvantages and present research efforts on impurity-based optimization of decision trees, as well as the study of new paradigms such as the use of Rényi Entropy, the utilization of the Tsallis Entropy, combinations of impurity functions, complexity-aware tree splits, and tailored impurity function augmentation, makes it clear that there are efforts underway to improve the accuracy, robustness, and comprehensibility, as well as the processing complexity, for the tasks of decision tree algorithms.**

***Keywords*—Decision trees; impurity functions; split criteria; decision tree optimization; literature review**

I. Introduction

Decision trees remain widely used in machine learning and data mining because they are interpretable, flexible, and effective for classification and prediction tasks [1][2][3]. Their practical value is especially evident in applications where transparent decision rules are needed to explain model behavior and support human decision-making.

However, the performance of a decision tree depends strongly on the impurity measure used to select splits [2][3]. When impurity is estimated poorly, the resulting tree may overfit the training data, become unnecessarily deep, and generalize poorly to unseen samples [4]. These limitations are more pronounced in complex, noisy, or high-dimensional datasets, where conventional split criteria may not capture the structure of the data effectively.

Decision trees, despite their seemingly simple model, possess the potential for representing and exploring the intricate interdependencies. In addition, decision trees constitute an understandable and easily comprehensible model, making them applicable for exploratory and descriptive data analysis [5]. Nevertheless, if the tree-growing algorithm ineffectively calculates the impurity of the nodes, it would most possibly lead to biased decision tree analysis. The effectiveness of tree growing as an algorithmic process would, therefore, depend on the choice of the impurity measure of the nodes that splits the potential splits. The values of the selected variables influence the point of splits and the corresponding split rule. Since the split rule becomes understandable, the decision tree provides an easily comprehensible representation of the data concerning the classes and the selected covariates. The model would express a tree structure, and the process of interpreting the tree would involve the nodes, branches, and terminal nodes of the tree [6].

II. Background and Significance

In the vast arena of data mining, decision tree algorithms are considered one of the basic tools, much admired for their high degree of interpretability and manipulability. Several decision tree algorithms have come into existence over the years, each having their respective algorithms belonging to CART, ID3, C4.5, CHAID, and then the Random Forest algorithm proposed by Breiman. This wide range of algorithms available, though, becomes one of the most significant challenges, as there always exists the challenge of choosing the best tree algorithm for optimal performance.

It needs to be accepted that the decision tree algorithms, especially for continuous variables, always face the potential challenge of an infinite number of splits and always tend to preserve the distribution of class probabilities, which often leads to the model overfitting. Moreover, the greediness factor always poses a significant challenge, as the basic motive for making the decisions always becomes the search for the highest degree of purity as the basic splitting criterion. These factors, specifically the challenges faced by the decision tree algorithms, like the potential challenge of an infinite number of splits and greediness, are already known and addressed by tree pruning and the addition of the factor of randomness during the development stage [7][8].

III. Research Motivation and Objectives

Even to date, the decision tree remains a basic classification technique in machine learning. Despite the many benefits of the existing classification algorithms based on the decision tree approach, ID3, C4.5, and CART classification algorithms have some limitations in predictive accuracy when applied to imbalanced, noisy, or high-dimensional datasets. One of the main demerits of the classification algorithms discussed above can be attributed to the need to rely on a fixed set of impurity criteria. These include Shannon entropy and the Gini index.

Current research developments in the context of impurity function design techniques—such as the application of general concepts of entropy, Rényi entropy, and the Tsallis entropy approach applied to RS-entropy—have shown promising results. There also exists a great scope of empirical research in the context of the application of improvements in the impurity function. At the same time, there also exists a research gap in terms of the complex relationships of impurity in trees. This

research aims to improve the accuracy of the decision tree classification technique and prevent overfitting by designing new impurity measures. These research objectives are:

- To investigate the inadequacy of the classic measures of impurities (e.g., Gini index, Shannon entropy) in the context of decision trees for different datasets.
- To develop new impurity criteria involving general measures of entropy and the consideration of parameters for flexible evaluation of splits.
- To test the proposed impurity measures empirically on a series of benchmark databases in order to see how well the measures perform in terms of accuracy in comparison to traditional decision tree algorithms.
- To test the effect of the improved impurity functions when applied in combination with the concepts of pruning, ensemble learning techniques, and feature selection.
- To establish the statistical significance and generalizability of the devised methods by performing validation techniques such as cross-validation and non-parametric tests.

Achieving the above objectives would ensure the advancement of the existing decision tree approach to one that provides higher accuracy and robustness without compromising the need for interpretability in trustworthy AI.

IV. Structure of the Literature Review

The literature review in this study is organized into a logical and progressive structure designed to contextualize the role of impurity functions in decision tree construction, highlight current scientific gaps, and synthesize contemporary research addressing these gaps. The structure is intentionally crafted to guide the reader from foundational knowledge toward advanced theoretical and empirical developments. It is organized into the following major components:

A. Foundations of Decision Tree Methodology

This section discusses the theoretical background of decision tree algorithms, their history, and their universality within the machine learning and data mining communities. This section explains the important attributes such as explainability, hierarchical representation, and flexibility and how the use of impurity measures is fundamental to the design of high-performance tree-based algorithms. This section helps the reader understand the challenges of continuous features, greedy splitting, and overfitting.

B. Classical Impurity Functions and Their Limitations

This section of the study evaluates commonly employed impurity estimates, such as the Gini Index, Shannon Entropy, and Error, along with their definitions, mathematical formulation, and theoretical characteristics. This section lays considerable emphasis on the performance problems associated with conventional impurity estimation algorithms, which are prone to difficulties such as imbalance, noise robustness, scalability, and the generation of complex decision trees.

C. Emerging and Alternative Impurity Measures

This section integrates the latest research efforts for improvement and new forms of impurity values, such as the use of Rényi entropy, Tsallis entropy, MECC, and combinations. The latest research on comparisons between such impurity measures and classical impurity measures, as well as the advantages derived from tunable parameters and probabilistic formulations, are discussed.

D. Decision Tree Performance Challenges

After the discussions on impurity functions, the following section looks into the persisting issues involved in the creation of decision trees, such as overfitting, sparseness, missing values, instability, and lack of generalizability. This section combines the findings regarding the role of impurity functions in the above issues and the gaps for the creation of tailored impurity functions.

E. Complementary Optimization Techniques

Impurity measures alone cannot completely address the shortcomings of decision trees. In this section, we review methods that can improve performance when used in conjunction with improved impurity measures. The main strategies are:

- Pruning techniques, including cost-complexity pruning, error-based pruning, and bootstrap out-of-bag pruning.
- Feature selection methods, namely filter and wrapper approaches.
- Ensembling techniques, including Bagging, Boosting, and hybrid models.

Literature suggests that the incorporation of advanced impurity functions coupled with these complementary techniques results in decision trees that are more stable and generalizable.

F. Tree Optimization: State-of-the-Art Developments

This section covers modern developments, including gradient-based learning of decision trees, differentiable tree architectures, splitting criteria driven by changes in distributions, and parallelized frameworks for building trees. It also provides an overview of research into weakly supervised anomaly detection, financial modeling, educational analytics, and explainable AI to understand real-world usage and effectiveness of state-of-the-art tree optimization techniques.

G. Identified Gaps and Research Opportunities

The final section synthesizes the literature to identify specific enduring gaps:

- The need for impurity functions that are fit for complex, heterogeneous data.
- Lack of adaptive mechanisms for impurity selection.
- Limited integration with explainability and interpretability tools.
- A lack of theoretical guarantees for new impurity formulations.

These gaps motivate the current study, which seeks to systematically investigate and appraise customized impurity functions that enhance accuracy, reduce overfitting, and improve computational efficiency in decision tree learning.

V. Decision Tree: An Overview

A decision tree is indeed one of the most powerful algorithms in supervised learning for both classification and regression tasks. It is still the most widely used algorithm today and can be visually represented as a flowchart, where each leaf will correspond to the output that is expected from the decisions made from the internal branches. The tree needs to be structured from a root down to the leaves by balancing readability and performance (see Fig. 1) [9]. It should be compact and use easily interpretable, stable variables with solid classification power as measured by its impurity. This is significant in keeping the tree concise for making sure predictive strength is maintained, especially in high-stake decisions where early trees should still accommodate some level of uncertainty [2].

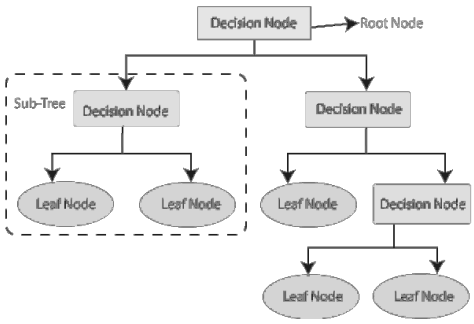


Fig. 1. Traditional decision tree [10].

A. Definition and Characteristics

Decision trees are generally known and accepted as basic techniques used in machine learning. The flexibility offered by decision trees is due to the ability to handle various attribute types, such as nominal, ordinal, interval, and continuously scaled attributes. In nominal attributes, the value is tested, and specific value information is used to transition to the next leaf [11]. In contrast, neighboring value ordering is used to establish a cut point, showing symmetry to the cut point, with the attribute value of the present case indicated by belonging to the region defined by the cut points. When the value of the region and corresponding cut points are divided into the region defined by the cut points, the attribute itself is said to be multidimensional. Varying the value of the region and the cut point could trigger a leaf transition, showing that the decision tree structure could vary by adjusting attribute values to the figure defined by quadruplets [12]. The ability to divide and conquer nominal and ordinal attributes enables the easy convergence of decision trees to attribute types. Further, using trees to continuously scaled attributes is enabled due to the structure of the leaf node, allowing the use of decision trees on problems involving interval, spatial, and continuously scaled attributes [13].

Forward pruning techniques, used to improve simplification models, are normally dependent on validation data, which is further divided into training and validation samples. These algorithms work to optimize decisions to obtain the lowest possible tree, which is guaranteed to keep validation samples having lower individual errors than training sets' probable errors. This is known as the cost-estimation approach, which expects methods used to estimate decision costs to avoid errors associated with performance, so as to derive an expansive and correct tree to predict and adapt to new and changing test/live situations.

B. Common Impurity Functions

In light of identifying Gini impurity and entropy as the most common impurity methods, the importance of the essential differences between their usages in decision tree building is discussed [14]. In comparison to more balanced datasets, Gini impurity is generally more biased toward the majority class, while entropy is generally more balanced and tends to generate smaller trees. In support of the comparative assessment, it has been found to work best on balanced datasets and class distributions [14]. This could explain why, under various datasets, Gini impurity has been found to outperform entropy on several instances [15][16]. When most instances are negligible, it has been observed to have less bias compared to entropy, which tends to reduce differences between impurities by biasing individuals to identify a more significant categorization ordering, albeit generating more suitable root node splittings [17]. In contexts when explainability is most paramount, the smaller-sized trees generated by Gini impurity could prove to be more beneficial, owing to their ease of understanding. Moreover, it is also more applicable to real-case problems due to high efficiency [18][19].

VI. Impurity Functions in Decision Trees

In decision tree models, nodes are represented by the attributes contained in the datasets, and the leaf node is used to define the class to which the datasets belong [20]. In discriminative learning, it is essential to identify the decision nodes by calculating the impurities, Eq. (1), present at the nodes used to divide the datasets. Given the fact that a node holds the impurities present in the datasets, the choice of input variables is important at the node during learning [21]. Various impurity measures have been developed to cope with input variable selection [21]. Impurities used to build decision trees include basic measures on errors, such as information entropy, the Gini index, and rule-based methods [22]. Using various impurities and a fixed probabilistic constraint along with learning on nodes, the idea of I-Optimal trees is developed based on the basic set of I-Optimal responses to probability, which aims to minimize the maximum non-negative relative entropy between the errors and the trees, and to develop insertable trees based on information functions and two-class tree optimization [23][24].

$$\text{impurity} = 1 - \text{purity} \tag{1}$$

Decision trees have found broad applications in data mining [25]. Decision trees have long been known to effectively and significantly provide easier methods to understand complicated procedures involving decision-making [26]. Moreover, the interpretable attribute of decision trees is unique to the algorithm,

adding to their popularity [27]. Compared to other models used to perform classification, decision trees are surprisingly more interpretable [28]. Decision trees have found broad applications ever since their creation and belong to a broad array of fields, including but not restricted to, bioinformatics, simplification of constraint programming problems, engineering, machine learning, modeling survival probability of trees, natural resource management, learning parity problems, and virtual screening, apart from applications to ensembles and forests [29][30]. Various extensions to the basic concept of decision trees have been developed, including but not restricted to, conic trees, conditional inference trees, decision stumps, frequent decision trees, interactive fuzzy decision trees, multigrain trees, multitrigger model trees, quantile random forests, totally corrective linear decision trees, and weight boosting [11][31].

A. Gini Impurity

The Gini impurity has several advantages. First, it significantly handles the issue of class imbalance. While it is ideal to have a balanced learning set, the Gini impurity has a tendency to lean towards the smaller classes, thus increasing their chances of reaching the top of the decision tree [32]. Moreover, it places the most frequent class at the top of the tree, having a greater influence than any other class. As a result, it quickly divides the large set into the upper nodes of the tree. Thus, the Gini impurity is often considered one of the most suitable impurity measures to be used in decision trees and random forests. The integrity of the tree model would significantly be affected if inappropriate measures were used [33]. In addition, the Gini impurity is also effective in reducing the complexity of the tree, and the efficiency of the tree-building algorithm is increased due to the reduced number of operations [34]. Moreover, it has been evident that the computational complexity of the Gini impurity is lower than information gain, and it is computationally efficient [35].

The Gini impurity measures the probability of randomly choosing an instance that would be assigned to a different class than it should. For a node t, it is defined as Eq. (2):

$$\mathrm{Gini}(t) = 1 - \sum_{j=1}^{k} p_j^2, \tag{2}$$

where, p_j is the proportion of instances of class j in node t and k is the number of classes [36]. The Gini impurity has value 0 when instances belonging to a node are maximally pure (all in a single class), and in the binary case it reaches its maximum of 0.5 when instances are equally distributed between the two classes, which corresponds to maximal impurity [37].

B. Entropy

Mathematically, the concept of entropy has been used to define the measure of Eq. (3) and Eq. (4) disorder or unusable energy present at the molecular level [38]. In contrast, the concept of entropy has been used to define the average amount of information, the measure of uncertainty associated with a source, and also the amount of randomness associated with a message. In essence, the more diversity, the more entropy [39]. Another concept, conditional entropy, is used to define the remaining uncertainty associated with the random variable, having knowledge of another variable, and is given by $H(Y|X) = S(I)$ [39]. There is a relationship between conditional entropy and entropy. In decision trees, the concept of information has been defined to mean the impurity associated with a dataset [40].

$$E(S) = -\sum_{x \in S} p(x) \log_2 p(x) \tag{3}$$

$$IG(S, A) = E(S) - \sum_{t \in T} p(t) E(t) = E(S) - E(S|A) \tag{4}$$

1) Rényi entropy: Current literature has further revived interest in using Rényi entropy, Eq. (5), to increase the correctness of decision trees [41]. This has been proven to outperform standard methods of measuring uncertainty and has prospects to tackle problems of unbalanced datasets. The impact of several Rényi measures on the efficiency of decision trees has also been analyzed, reflecting the ability to measure uncertainty associated with datasets to improve the accuracy of models [42]. Studies have also attempted to measure Rényi uncertainty along with other measures to increase decision tree efficiency [43].

$$R(D) = \frac{1}{1-\gamma} \log \left(\sum_{i=1}^{k} p_i^{\gamma} \right) \tag{5}$$

2) Tsallis entropy: When used as an impurity measure, Tsallis entropy, Eq. (6), has been proven to increase the efficacy of decision trees, as the inclusion of unconventional measures of impurity, including Tsallis entropy, has led to optimal results in decision tree models [44]. The importance of understanding the potential impact of Tsallis entropy on decision tree performance should not go unnoticed. Moreover, the inclusion of Tsallis entropy has led to significant improvements in the accuracy of decision trees. Cumulatively, the literature suggests that the inclusion of Tsallis entropy could prove to be a worthy tool to improve decision tree performance [41].

$$S_q = \frac{1 - \sum_{i=1}^{k} p_i^q}{q - 1} \tag{6}$$

C. Classification Error

The impurity measures under consideration demonstrate the importance of weights used when determining impurities [45]. The classification error function, Eq. (7), is used as an impurity measure on the terminal set and is a measure used to evaluate the splitting attribute, which depends on the selection attribute used to decide on splitting [46]. The classification error function is universally suited to a number of classes and is known by the name Gini impurity coefficient when used on two classes, and it is one of the impurity measures used in the empirical field. When achieving accuracy is assessed, a high-value item could incur more costs than a lower-value item [47]. The relevance of understanding measures of cost or loss is critical to accurately compute splittings and variables [48].

The aim of the present review is to demonstrate the potential role of the use of ranging functions on the flexibility offered by decision trees to social scientific studies [49]. In some situations, it is also possible to utilize a boosting algorithm with closely similar statistical power to handle various MC matrices [50].

$$E(D) = 1 - \max_i(p_i) \tag{7}$$

VII. Decision Tree Performance Challenges

Decision tree induction involves mainly three broad phases: formation of the tree structure, pruning, and leaf node assignment. Formation of an efficient decision tree is very important, and it is always desirable to have a decision tree with similar class distributions of the child nodes [12]. Nevertheless, it is often difficult to satisfy such a restriction when working on practical problems. In addition, it is more complex to develop a homogeneously structured decision tree with morphological information than to generate an effectively shaped decision tree [29].

Decision trees are a widely used technique for classification and prediction. The strength behind decision trees lies in their simplicity, usefulness, and ability to capture relationships with a fewer number of restrictions on their structure. This technique is helpful when it is required to work on moderate to large-sized datasets and involves very little input information on specifying the problem. It is, however, important to keep in mind that decision trees are vulnerable to noise present within the input, which often causes the creation of a large number of rules [51].

A. Overfitting

As data scientists and machine learning experts, researchers have the power to set criteria to govern tree models to avoid overfitting. This is possible by incorporating the number of training samples (n) and the number of relevant variables to predict the target variable (m), such that n is at least equal to $2m$ [52]. The number representing relevant variables is used either in the leading section of the data divide or by adding up all the conditions used to classify each training sample to the target category. This enables each category to hold distinct activated conditions to facilitate the validity of inferential modeling. Although simplistic, it is very effective to control the structure and number of relevant input variables to grow the decision tree modeling approach. Other techniques involving the effective utilization of weighted and optimized systems to identify relevant attributes and unique amounts of data to correctly categorize are also capable of increasing the efficacy of the aforementioned rule to $n \geqslant 2m$ [53].

The cause of overfitting can therefore be said to result from the inclusion of noise and unnecessary attributes, which the tree is trying to capture and employ to classify or explain the training data. The overfitting of the data is not sudden during the attribute selection procedure but starts with the selection of the optimal attribute and the inclusion of less optimal attributes to the tree nodes. The tree continues to grow until a considerable amount of irrelevance and interference is injected into it. The problem occurs because the training data have been optimized solely to suit the specific data the tree is growing on and may result in incorrect classification during the presentation of new instances [54]. Beginning with decision trees, overfitting is a basic flaw associated with them owing to their ability to learn and generalize data but not accurately enough. Noisy data during training are also associated with this flaw. When the attributes of the tree grow to complexity with too many nodes compared to the number of training instances, it is said to experience overfitting, marked by too much unnecessary modeling [52].

B. Handling Missing Values

Decision trees are commonly used to classify instances within a specific domain defined by a problem. The instances are divided into several sub-domains, and a model is built based on each domain and a set of attributes. This is conducted recursively. Several other learning methods are also found, which contain nodes not connected to decision trees. These methods are mainly used to improve the induction of decision trees [55].

Decision trees with missing information can also be handled using measures of impurities, which determine how to treat the missing information. Missing information may require filling using the most frequent value, and another measure of impurities may also be used to measure the influence of the missing information on the decision tree. When attributes are continuous, the average is commonly used, and when discrete, the mode is used. After structuring information and classes associated with information, the algorithm determines the best splitting and modes.

On the other hand, Quinlan developed the C4.5 algorithm to build decision trees, which works effectively with attributes that require either continuous or discrete information. It divides instances into various areas defined by a test and assigns instances to paths showing maximum information gain. In the meanwhile, Fernando et al. have made improvements to develop more efficient training and increase the efficiency of neural learning by reducing the influence of missing information [56]. Some authors also go into detail on how missing data are handled during splitting when constructing decision trees, while others consider it another value to take into consideration along with other attribute values. As a result, the diversity discussed above encompasses a wide review of the subject to guarantee readers are knowledgeable on the various perspectives on the issue [57].

VIII. Enhancing Decision Tree Performance

Current algorithmic developments involving decision tree impurity are an interesting solution, although not without restrictions [58]. Moreover, the existence of impurities is not convex, thereby increasing the complexity of reaching an optimal solution, and it is not guaranteed to converge within a fixed time span dependent on the input attributes [59]. Thus, it has also been realized that, to a great extent, the rigid measure of impurity is often settled with the measure of minimum quality, with the hope of achieving optimal models with impurities below a certain threshold. Unfortunately, such relaxation of impurity measures has further triggered unwanted occurrences of overfitting decisions on the decision trees [60].

Decision trees, although having high capabilities on classification problems, are also faced with demerits, largely due to the presence of noisy information in the datasets. Most learning models are also known to be highly affected by impurities, and more so when it is regarding class imbalance [28]. Although it has been realized to a great extent that having robust positions against impurities and enough data regarding impurities to learn is important, impurity is not invariant. If it is considered to a large extent, it is known to negatively affect learning models, and if the impurities are considered to a lesser extent, it often has to consider a rule regarding the pure attributes of the minority class and is further embedded into the decision trees. Under such conditions, it has been realized to largely overfit the trees, correctly categorizing instances in the minority classes and largely misclassifying instances in the majority classes [61].

A. Pruning Techniques

There is a possibility of overfitting during the creation of a decision tree, which may result in the creation of a very large tree [62]. Although the aim is to develop an apt representation of data, it often tends to remember the data and perform very well on it. This problem can be mitigated through pruning. Pruning the tree can be achieved by cost-based pruning using the resubstitution error, the estimate of the error, or by methods such as the bootstrap out-of-bag error, which serves as a validation technique and is very useful, particularly when reducing the time complexity [63]. The bootstrap out-of-bag error can also act as a third option that combines these techniques and serves as a validation procedure, and is very helpful, particularly when reducing the time complexity.

The entire process of pruning allows the creation of an unbiased terminal tree, which provides efficient access to the basic structure of the data, namely the relation between impurities and the decision tree [64]. By using metrics such as the Gini index and entropy, along with validation and techniques such as the random forest, it is possible to increase robustness, which allows the creation of a better and more universally applicable model [65]. Decision tree methods are used owing to their immediacy and ability to divide observations into separate subsets, which results in them becoming an extremely efficient tool and very difficult to reproduce and verify by other techniques such as multivariate methods [27]. Decision trees along with impurities are flexible methods and can handle nominal and continuous independent variables [4]. The technique, which is similar to a set of binary tests, is able to identify optimal cutting points on the independent variables to increase the purity of the groups, which is very helpful to behavioral analysts and market experts, and is able to provide very rapid characterization and understanding of the data. Hence, it is able to act as an extremely efficient description tool, identifying the region of interest surrounding the most basic subsets of various variables [66].

B. Feature Selection Techniques

Input feature priority techniques imply ranking input features based on relevance to the target task. This is helpful to identify the optimal number and priority of input features to retain. Input feature selection is divided into two broad categories: filter and wrapper techniques. In filter techniques, importance is ranked, and measures such as the t-statistic and F-statistic are used. These techniques can select subsets of top-k features and are used during the first stages of data preprocessing. There are various input feature selection techniques, and some of the most important ones are discussed below.

In contrast, wrapper methods utilize a search procedure along with a model to assess the performance of various subsets of features [67]. Features are selected and dropped based on their ability to improve or hinder the performance of the model [68]. In contrast to filters, the wrapper technique can select a feature that is important to the model yet not directly associated with the target, and drop a feature that is less important but directly associated with the target [3]. As such, the wrapper technique often outperforms filters [69].

The objective of the feature selection procedure is to discover a set of attributes within the input features to enable the creation of a predictive model with enhanced accuracy, which can develop a more efficient predictive model by learning the most distinguishing features while avoiding unwanted characteristics [70]. Several advantages are associated with feature selection, including the creation of more interpretable models, the ability to develop models with faster speeds and smaller sizes, the elimination of variables highly associated with the target, and the attainment of negligible value. There are two types of feature selection methods [5, 71].

C. Ensemble Methods Techniques

In the decision trees literature, an ensemble technique combines N models (developed by using the same input learning algorithm) to attain a unique prediction for every instance within the test data [72]. There are two basic methods to design an ensembling technique of decision trees [73]. Bagging involves creating an ensemble by randomly sampling, with replacement, elements within the training example and using the sample to develop the respective model [74]. Boosting techniques are based on an iterative technique wherein models are developed on the entire training example, and weights are finally given to poorly predicted instances (see Fig. 2) [75]. These poorly predicted instances also display a greater tendency to be selected into the training example of the next models [76][77]. These methods are further coupled with decision trees to generate an ensembling technique of models [78][79].

1) Bagging: When it is used correctly, involving decision trees as base models, bagging can generate very good predictive models. However, it should never be forgotten that bagging works mainly to decrease variance, which is why it is most beneficial when used on high-variance models. Another important factor is to include homogeneous and robust base models. When base models are not satisfactory, it is possible to improve performance by differently initializing base models randomly and using differently defined discretization methods, such as divide techniques and Vapnik methods [80]. While it is true that, when used correctly, bagging and resampling methods can often involve more active and complex processes than other techniques used to generate ensembles, it is also very important to understand and appreciate the strength of

bagging, which is to accurately and efficiently develop predictive models [81].

The basic bagging algorithms require a robust learning algorithm, which is true when bagging is used with decision trees [11]. When used on decision trees, bagging produces a mixture model where each individual model is obtained using a bootstrap sample of the training instances, and the training involves all instances and variables in the training set. The predictions of the individual models (decision trees) are averaged in regression and combined via voting in classification [76]. The best possible results are achieved when the training is carried out on noisy data and data with a high dimensionality. Bagging is most effective when it decreases the variance of the base learner when it is unstable.

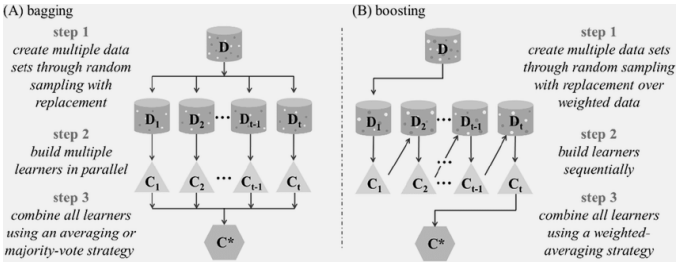


Fig. 2. Bagging and boosting tree-based ensemble [10].

2) Boosting: XGBoost, which is a very reliable and highly researched boosting algorithm, is very similar to residual-fitting boosting but is distinguished by adding a specific weight to each classifier based on the accuracy rate. The weights are incorporated into the overall classifier, which makes XGBoost a forward stage-wise fitting algorithm [82]. Unlike traditional methods, XGBoost gives more weight to instances that were classified inaccurately to influence the learning process of new classifiers. The concept of boosting has the ability to improve a weak solution to a strong solution [6].

Boosting techniques present the problem of optimization by working on iterative improvements of specific loss functions by training on instances with more importance. The next classifier is dependent on the most important instances, which are determined by the distribution defined by the earlier classifier. These techniques, such as AdaBoost and RankBoost, increase the importance of instances by modifying the distribution after each classifier. Other boosting techniques, such as BrownBoost, work differently by approximating the loss function without changing the distribution after each classifier [73][79].

IX. Related Work

Mao et al. [83] present a new decision tree classification algorithm based on a two-term RS-entropy concept, aiming to improve the flexibility and accuracy of models. The authors highlight the insufficiencies of existing decision tree approaches such as ID3, C4.5, and CART, which rely on rigid splitting rules defined by Shannon entropy and the Gini index. RS-entropy, a generalized concept encompassing Shannon, Rényi, Tsallis, and r-type entropies, allows dynamic splitting rules defined by adjustable parameters. Furthermore, the authors enhance the two-term information formula by incorporating penalty factors to account for attribute complexity, resulting in the RSEIM algorithm. Empirical experiments on eight UCI datasets show that RSE and RSEIM greatly outperform traditional algorithms in classification accuracy without increasing model complexity. Genetic algorithm–based parameter optimization further strengthens RSEIM's robustness and resistance to overfitting. Pairwise comparisons and Nemenyi tests confirm the statistical superiority of RSEIM.

He et al. [84] propose a dynamic Bayesian node-splitting algorithm to improve the performance of the random forest algorithm by overcoming inefficiencies in traditional splitting methods. The algorithm dynamically calculates posterior probability distributions during the splitting process, enabling more accurate identification of optimal split points. The authors also introduce a dynamic thresholding mechanism weighted by misclassification errors to prevent overfitting. Experiments conducted across multiple UCI datasets show improvements in accuracy, robustness, and computational efficiency compared to classical random forest models.

In the work of Ferreira et al. [85], decision tree models are used to support clinical decision-making regarding impaired physical mobility in trauma patients. Using the C4.5 algorithm, the authors identify eight essential nursing diagnoses and construct classification trees to determine key predictors such as age, length of hospital stay, and injury severity score. The results demonstrate that decision trees can effectively guide clinicians in prioritizing patient care and integrating data-driven assessments into trauma management.

Ciftcioglu et al. [86] introduce a predictive model for estimating shear strength in reinforced concrete beams, optimizing decision trees using the Lévy Flight algorithm to overcome the limitations of traditional machine learning approaches. Using an extensive dataset featuring numerous structural variables, the authors compare classical decision tree models with Lévy Flight–enhanced variants. Results show that the optimized models produce superior accuracy, robustness, and adaptability while reducing overfitting. Metrics such as RMSE, MAE, and R^2 confirm that the new method significantly outperforms traditional decision trees.

Li et al. [87] present an extensive analysis of a decision support system (DSS) for vending machines using machine learning techniques, including an improved C4.5 decision tree algorithm and neural networks. The modified C4.5 algorithm reduces runtime and slightly improves accuracy relative to the original. Using a backpropagation neural network (BPNN), the authors show strong alignment between predicted and real sales trends. Reinforcement learning is also incorporated to further enhance the DSS's efficiency and forecasting accuracy.

New splitting rules for Gradient Boosting Decision Trees (GBDT) are proposed by DeLise [88] to address out-of-distribution (OOD) generalization—a major challenge due to era-dependent distribution shifts. Unlike Empirical Risk Minimization (ERM), which assumes i.i.d. data, the new rules leverage cross-era information to build more robust models. Theoretical analysis and experiments using simulated datasets and real-world financial market data (e.g., Numerai) show improvements over classical GBDT.

Marton et al. [89] introduce GradTree, a technique for learning hard, axis-aligned decision trees using gradient de-

scent rather than traditional greedy methods. By employing backpropagation and the straight-through estimator, GradTree jointly learns all tree weights, producing enhanced accuracy and generalization performance in binary and multiclass tasks. Experimental results demonstrate that GradTree surpasses existing algorithms while supporting external loss functions and dynamic splitting.

Zeng [20] examine fundamental assumptions about impurity functions in decision trees, focusing on non-negativity and concavity. The authors show that impurity functions are not inherently non-negative and analyze conditions where impurity decreases may be harmful. They further demonstrate that concavity guarantees non-negative impurity decreases in multiway splits. Experiments on the German credit dataset show that a linear hybrid of the Gini index and entropy outperforms each impurity measure individually.

In their study, Arbiv et al. [90] explore human-centered optimizations for decision trees to improve interpretability. The authors evaluate design modifications and algorithmic enhancements aimed at reducing the gap between complex data structures and user understanding. Their findings show that optimally designed trees significantly improve human decision-making by presenting information in a more comprehensible format.

Loreti and Visani [91] assess parallelization strategies for building decision trees in the GLEAMS explainability framework, which struggles with scalability. By applying distributed computing through the Ray platform, they compare multiple parallel methods and identify performance strengths and weaknesses. The results highlight improvements in efficiency and scalability in alignment with modern explainability demands.

Finke et al. [92] investigate tree-based methods for weakly supervised anomaly detection at the Large Hadron Collider (LHC). Their work demonstrates that Boosted Decision Trees (BDTs) outperform deep neural networks in training efficiency and robustness to noise. Using gradient boosting and ensemble techniques, the authors show that BDTs can effectively operate in high-dimensional spaces with many uninformative variables, improving model-independent searches for new physics phenomena.

Baena-Rojas et al. [93] present a hybrid approach to Multi-Criteria Decision-Making (MCDM) to identify international markets for small and medium-scale coffee roasting businesses. The hybrid approach combines and balances the benefits of quantitative and qualitative techniques to improve the efficiency and viability of International Market Selection (IMS). A case study involving 18 Colombian SMEs producing coffee shows that cultural variables and techniques improve the efficiency of the MCDM solution. The most promising countries are the US, the Netherlands, and South Korea, demonstrating the feasibility of the proposed solution.

Ujwal and Malik [94] suggest a system called PerformanceX, which utilizes Educational Data Mining (EDM) to enhance performance and minimize dropout ratios. The authors present a selective algorithm involving filter and wrapper methods to select important attributes based on student datasets. Using machine learning classifiers, including a fusion voting classifier, the system achieves an accuracy of 99.41%, with the best result obtained by the random forest classifier. The work clearly emphasizes the relevance of structuring student performance on effective attributes to improve learning and career advancement.

Bouke et al. [18] introduce the BukaGini algorithm, a new feature selection algorithm suitable for machine learning. The algorithm leverages the concept of the Gini impurity index to improve the evaluation of interactions between features, including linear and nonlinear associations. Comparisons conducted on various datasets, such as High School Students' Performance, Cancer Data, Spambase, and UNSW-NB15, indicate that BukaGini generally attains greater accuracy improvements than conventional Gini index methods, ranging between 0.32% and 2.50%.

Pathan and Sharma [95] analyze improvements to the decision tree algorithm to increase classification accuracy and minimize tree size. Their discussion of preprocessing techniques focuses on handling missing information and uncertainty. The proposed decision tree algorithm, optimized using the Gini impurity, is more efficient than regular decision trees based on criteria such as precision, recall, F-score, and accuracy.

Subramani and Muruganantharaj [96] propose the EnhancedTree+ algorithm, an advanced decision tree classifier designed to enhance accuracy, interpretability, and robustness when dealing with complex datasets. Although decision tree algorithms are easy to implement, they suffer from overfitting and loss of interpretability when handling complex data. EnhancedTree+ addresses these problems by incorporating smarter splitting rules and pruning. Extensive experimentation on the Titanic dataset shows that EnhancedTree+ outperforms standard decision tree classifiers and competes favorably with models such as random forests and gradient boosting. The experiments clearly indicate improvements in accuracy, precision, recall, and F1-score, implying strong practical viability.

Asif et al. [97] address global health issues related to cardiovascular diseases by presenting a machine learning solution involving preprocessing, hyperparameter tuning, and ensemble learning techniques. The authors combine three Kaggle datasets with similar attributes to construct an extensive dataset. Using an Extra Trees classifier, normalization, and grid search cross-validation, the proposed solution achieves an effectiveness of 98.15%. The results suggest strong potential to correctly identify cardiovascular disease, which may greatly aid in prevention, diagnosis, and treatment, thereby reducing morbidity and mortality.

The work by Cui et al. [98] targets the robustness of Gradient Boosted Decision Trees (GBDT) and offers a technique to improve it by employing one-hot encoding and regularization. Although efficient, GBDT models are vulnerable to slight changes in input data. By converting the models into linear form using one-hot encoding and applying L_1 and L_2 regularization, substantial improvements in robustness are achieved. The study shows that regularization methods help alleviate vulnerabilities to data perturbations, ultimately improving efficiency.

Custode and Iacca [99] explore the integration of evolutionary algorithms and Q-learning to provide interpretable decision trees for reinforcement learning (RL). Their work tackles problems such as premature convergence to local optima by adopting a two-level optimizer. Experimental analysis

conducted on classical RL tasks (CartPole-v1, MountainCar-v0, and LunarLander-v2) and on a pandemic control domain clearly shows the competitive efficiency of the proposed solution for both interpretable and non-interpretable models. The results demonstrate that interpretable models can match or even improve upon the performance of non-interpretable ones, even in high-stakes domains such as pandemic control.

Riansyah et al. [100] publish a study on improving the accuracy of the C5.0 decision tree algorithm for data mining classification using adaptive boosting (AdaBoost). They highlight several flaws of the C5.0 algorithm, including susceptibility to overfitting and accuracy degradation when dealing with unbalanced data. These flaws are mitigated by AdaBoost, which adjusts the weights of incorrectly classified samples. In a case study involving hepatitis data, accuracy is increased from 80.58% to 82.98%, while the classification error rate is significantly reduced. This shows AdaBoost's ability to transform a weak classifier into a strong one through reweighting of training samples.

In their work, B et al. [101] present Permutation Decision Trees (PDTs), an extension of classical decision trees that uses the Effort To Compress (ETC) complexity measure—interpreted as an impurity measure—to enable permutations of instances. Compared to Shannon entropy and Gini impurity, ETC captures the importance of instance ordering, making PDTs sensitive to permutations. The authors also propose Permutation Bagging, an implementation of PDTs that does not rely on stochastic sampling but instead uses random feature selection. Although the approach has promising applications, comparisons with standard random forests clearly show the superiority of Permutation Bagging and Permutation Decision Forests, particularly when handling critical data instances and dependencies.

Awad and Fraihat [102] study the efficacy of recursive feature elimination with cross-validation using decision tree models (DT-RFECV) for feature selection in intrusion detection systems (NIDSs). Employing the UNSW-NB15 dataset with 42 features, they improve accuracy while reducing the number of features. The findings reveal that using the optimal 15 features obtained by DT-RFECV yields a binary classification accuracy of 95.30%, comparable to that achieved using the full feature set. This work maintains high accuracy while speeding up detection and reducing storage requirements by 65% through feature selection.

Zhang and Gionis [16] propose a new algorithm for decision tree induction that incorporates complexity guarantees during learning. Although classic algorithms such as C4.5 and CART are effective in terms of accuracy, they offer no guarantees regarding the complexity of the constructed trees and may thus produce poorly interpretable models. The new algorithm improves on these approaches by adding a greedy heuristic that balances tree size, cost, and impurity, yielding logarithmic guarantees on complexity. As a result, it produces trees that are both interpretable and accurate. The paper represents an important contribution by explicitly addressing the trade-off between accuracy and interpretability.

Roshanski et al. [103] propose a new algorithm, FACET, designed to ensure that decision trees remain compact while preserving modeling power. Although decision trees often provide a simple and robust modeling framework, their growth can lead to overly large and complex models. FACET seeks to maintain compactness by using automated modeling principles and feature extraction to construct concise tree representations. Experimental analysis on 82 different datasets is conducted to verify the efficiency and effectiveness of the proposed technique.

Bodine and Hochbaum [104] present the Max-Cut Decision Tree with modified PCA, published in SN Computer Science, with the aim of optimizing the decision tree algorithm. Their work remedies problems associated with standard CART Gini trees by incorporating two major innovations. First, Max-Cut is designed to maximize separation between class instances, making it less prone to overfitting. Second, the integration of Node Means PCA dynamically adapts dimensions at each node, reducing running time and improving efficiency. Experimental validation across multiple datasets, including CIFAR-100, shows considerable performance improvements.

Horváth et al. [105] propose a new method for adding robustness to the certification of decision-stump ensembles via deterministic smoothing. Although tree-based models are widely used in high-risk domains such as healthcare and finance, their robustness certification lags behind that of neural networks. The authors introduce deterministic robustness certificates for ensembles with numerical and categorical attributes. Their method uses dynamic programming to compute exact output distributions under randomization and provides an MLE-optimal learner for robust stump-based boosters. Experiments demonstrate improved efficiency and accuracy compared to existing robustness-certification approaches for tree models.

Li and Chen [106] address the shortcomings of conventional univariate decision trees (UDTs) by introducing a new splitting rule based on Geometric Mean Metric Learning (GMML). They propose the diagonalized GMMLDT, which accelerates tree construction by quickly identifying the most discriminative splitting feature while preserving explainability. Experiments on standard benchmark datasets show that both dGMMLDT and its multivariable extension, dGMMLMDT, achieve comparable or superior accuracy with up to a ten-fold speedup relative to full GMML.

Lee et al. [33] propose a decision tree algorithm that improves classification accuracy by incorporating class distances. The method modifies the traditional Gini index using a combination of Rao and Stirling distances, producing more clearly separated class boundaries during splitting. Testing on multiple datasets—including contraceptive method choice, car evaluation, and yeast—demonstrates that the new algorithm consistently outperforms the traditional Gini index.

Manzali et al. [107] propose Enhancing Weak Nodes in Decision Trees (EWNDT), an approach designed to improve decision tree efficiency by addressing weak nodes using data augmentation. The algorithm augments data associated with weak nodes by artificially grouping neighboring nodes and re-evaluating candidate splits, thereby increasing model robustness. Experiments on UCI repository datasets show that EWNDT outperforms conventional pruning-based techniques.

Agarwal et al. [108] propose Hierarchical Shrinkage (HS), a post-processing regularization technique for decision trees

and random forests. HS shrinks predictions toward the mean of ancestor nodes through a single regularization parameter while preserving the tree structure. Experiments across varied datasets demonstrate that HS significantly improves predictive performance and interpretability. The authors also link their method to ridge regression applied to decision stumps, providing theoretical grounding.

Arabameri et al. [81] assess decision tree-based ensemble machine learning algorithms for landslide susceptibility mapping in the Taleghan–Alamut Basin, Iran. The study employs Credal Decision Tree (CDT) models integrated with Bagging, Multiboost, and Subspace algorithms using eighteen conditioning factors. Model performance is evaluated using ROC curves and k-fold cross-validation. Results show that CDT-Multiboost achieves the highest performance, with an average AUC of 0.993, indicating strong feasibility for generating reliable landslide-susceptibility maps to mitigate future hazards.

X. Methodology

This study follows a structured literature-review methodology rather than an original empirical design. The aim is to synthesize recent research on impurity measures used in decision-tree learning and to compare how the reviewed studies report performance, robustness, and interpretability.

The review methodology is organized into three phases:

- **Literature identification:** Relevant studies published between 2020 and 2025 were collected from the literature cited in this study, with emphasis on work addressing classical impurity measures, generalized entropy-based criteria, hybrid split functions, pruning strategies, feature selection, and ensemble-based decision-tree optimization.
- **Screening and categorization:** The selected studies were grouped according to their primary contribution, including impurity-function design, performance optimization, overfitting control, interpretability, and application domain. This step enabled consistent comparison across methods and research themes.
- **Comparative synthesis:** Findings were synthesized qualitatively by comparing the reported outcomes of the reviewed papers, including accuracy trends, robustness to class imbalance and noise, model complexity, and computational considerations. No new experiments were conducted in this study.

This methodology aligns the study with its actual scope as a survey and provides a transparent basis for the discussion and comparison presented in later sections.

XI. Discussion

The reviewed literature shows a clear evolution in the role of impurity functions in decision-tree construction. While classical approaches such as the Gini index and Shannon entropy remain dominant for their simplicity and computational efficiency, they exhibit significant limitations when applied to modern datasets characterized by imbalance, high dimensionality, noise, and non-linear interactions. As documented throughout the study, these drawbacks manifest through biased splits, deeper and less interpretable trees, and an increased tendency toward overfitting—an issue exacerbated by the greedy, locally optimal nature of standard tree-building algorithms.

Recent studies explore alternative impurity measures—such as Rényi entropy, Tsallis entropy, cost-sensitive error functions, MECC, and ETC-based measures—to more accurately quantify uncertainty in multi-class, imbalanced, or temporally sensitive datasets. These functions provide improved modeling flexibility by introducing tunable parameters, non-linear weighting schemes, or domain-specific structural priors. Importantly, emerging work demonstrates that combining impurity measures (for example, linear mixtures of Gini and entropy) can outperform individual functions for specific classes of problems.

Beyond impurity functions themselves, the literature underscores the importance of complementary techniques such as pruning, ensemble learning, feature selection, and error-based optimization to mitigate overfitting and excessive tree growth. Studies show that when decision trees employ improved impurity functions in conjunction with these techniques, the resulting models consistently outperform baseline CART and C4.5 variants in terms of accuracy, robustness, computational cost, and interpretability.

XII. Results and Comparison

This section summarizes the findings reported in the reviewed literature rather than presenting original experimental results. The comparison focuses on four recurring dimensions discussed across the surveyed studies:

- **Predictive performance:** Reported changes in accuracy, precision, recall, and F1-score.
- **Model complexity:** Tree depth, number of nodes, and resulting interpretability.
- **Robustness and overfitting:** Sensitivity to noise, class imbalance, and training-data variation.
- **Computational efficiency:** Reported training cost and scalability with larger datasets.

As summarized in Table I, the reviewed studies generally indicate that adaptive impurity measures and impurity-aware optimization strategies can improve decision tree behavior relative to classical split criteria, especially when combined with pruning, feature selection, or ensemble learning. These findings should be interpreted as a synthesis of published evidence, not as new benchmark results produced by the present study.

XIII. Conclusion

This literature survey demonstrates that impurity functions play a foundational role in shaping the performance, interpretability, and computational properties of decision-tree models. Traditional impurity measures, while effective for structured and balanced datasets, often fall short in modern applications characterized by complexity, heterogeneity, and noise. A new generation of impurity functions—incorporating tunable parameters, non-linear weighting, temporal sensitivity, or structural complexity guarantees—has shown substantial empirical and theoretical promise.

TABLE I. CONDENSED SUMMARY OF REVIEWED DECISION TREE RESEARCH (2020–2025)

ID	Authors / Year	Method / Algorithm	Key Contributions	Datasets	Results
1	Mao (2025)	RSEIM / RS-Entropy	Generalized entropy-based DT with tunable splitting	8 UCI	Higher accuracy vs CART/C4.5
2	He (2025)	Dynamic Bayesian Splits	Posterior-probability split optimization	UCI	Faster + more robust RF
3	Ferreira (2025)	C4.5 Clinical DT	Mobility prediction in trauma patients	Clinical	Better diagnostic insights
4	Ciftcioglu (2025)	DT + Lévy Flight	Optimization for nonlinear engineering data	Concrete beams	Improved RMSE, MAE, R^2
5	Li (2024)	Modified C4.5 DSS	Faster DT for vending machine DSS	DSS data	Higher accuracy, reduced runtime
6	Delise (2024)	Era-aware GBDT	Robust OOD generalization under distribution shifts	Numerai	Strong OOD performance
7	Marton (2024)	GradTree	Full-tree gradient-descent optimization	Benchmarks	Higher accuracy vs greedy DT
8	Zeng (2024)	Impurity Theory	Clarifies impurity properties; beneficial mixtures	German Credit	Mixed measures outperform baseline
9	Loreti (2024)	Parallel DTs (Ray)	Scalable parallelization for explainability	GLEAMS	Faster DT building
10	Finke (2024)	BDTs for Weak Supervision	Noise-robust anomaly detection	LHC	Better than deep nets
11	Baena-Rojas (2023)	Hybrid MCDM Tree	IMS ranking for SMEs	18 SMEs	Identified US, NL, KR as best markets
12	KVG (2023)	PerformanceX EDM	Attribute selection + DT	Student data	99.41% accuracy
13	Bouke-Bukagini (2023)	BukaGini	Interaction-aware impurity	UNSW, UCI	+0.32–2.50% accuracy
14	Pathan (2023)	Improved Gini DT	Missing-data robust preprocessing	UCI	Higher precision/recall
15	Subramani (2023)	EnhancedTree+	Improved splitting + pruning	Titanic	Beats DT/RF/GBDT
16	Asif (2023)	Ensemble Health Trees	Extra Trees + preprocessing	Kaggle heart	98.15% accuracy
17	Cui (2023)	Robust GBDT	L1/L2 regularized GBDT	Real + synthetic	More robust predictions
18	Custode (2023)	EVO + Q-Learning Trees	Interpretable RL trees	RL benchmarks	Competitive with deep RL
19	Riansyah (2023)	C5.0 + AdaBoost	Boosting for imbalance	Hepatitis	Accuracy ↑ from 80.6→83%
20	Bodin (2022)	Max-Cut DT + PCA	Class-separating splits + node PCA	CIFAR-100	49% better accuracy, 94% faster
21	Horváth (2022)	Robust Stumps	DP-based exact robustness certification	Mixed datasets	Better certified robustness
22	Li (2022)	GMMLDT	Metric-learning-based DT splitting	Benchmarks	Up to 10× speedup
23	Lee (2022)	Distance-based Gini	Rao–Stirling distances in impurity	UCI	Higher accuracy
24	Manzali (2022)	EWNDT	Weak-node data augmentation	UCI	Superior to pruning
25	Arabameri (2022)	CDT + Multiboost	Uncertainty-aware hazard mapping	Landslides	AUC = 0.993

Moreover, the synergy between enhanced impurity functions and complementary techniques such as pruning, ensemble learning, feature selection, and hybrid modeling emerges as a critical factor in achieving high-performing and robust tree-based models. These developments indicate a decisive shift toward impurity-adaptive and context-aware decision-tree learning frameworks.

This study thus underscores the importance of rethinking impurity measurement in decision-tree design to meet the demands of increasingly challenging data environments and motivates future work toward systematically designing, analyzing, and benchmarking customized impurity functions.

ACKNOWLEDGMENT

The project was funded by the KAU Endowment (WAQF) at the King Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for their technical and financial support.

REFERENCES

[1] M. J. Kochenderfer, T. A. Wheeler, and K. H. Wray, *Algorithms for decision making*. The MIT Press.

[2] G. Pappalardo, S. Cafiso, A. Di Graziano, and A. Severino, "Decision tree method to analyze the performance of lane support systems," vol. 13, no. 2, p. 846, number: 2 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2071-1050/13/2/846

[3] T. Zhang, H. Wang, J. Chen, and E. He, "Detecting unfavorable driving states in electroencephalography based on a PCA sample entropy feature and multiple classification algorithms," vol. 22, no. 11, p. 1248. [Online]. Available: https://www.mdpi.com/1099-4300/22/11/1248

[4] E. Scornet, "Trees, forests, and impurity-based variable importance." [Online]. Available: http://arxiv.org/abs/2001.04295

[5] W. Liu, H. Fan, and M. Xia, "Credit scoring based on tree-enhanced gradient boosting decision trees," vol. 189, p. 116034. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421013798

[6] M. A. Carreira-Perpinan and A. Zharmagambetov, "Ensembles of bagged TAO trees consistently improve over random forests, AdaBoost and gradient boosting," in *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*. ACM, pp. 35–46. [Online]. Available: https://dl.acm.org/doi/10.1145/3412815.3416882

[7] A. Zharmagambetov, S. S. Hada, M. Gabidolla, and M. A. Carreira-Perpinan, "Non-greedy algorithms for decision tree optimization: An experimental comparison," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/9533597/

[8] T. Thomas, A. P. Vijayaraghavan, and S. Emmanuel, *Applications of Decision Trees*. Springer Singapore, pp. 157–184. [Online]. Available: http://link.springer.com/10.1007/978-981-15-1706-8_9

[9] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," vol. 2, no. 1, pp. 20–28. [Online]. Available: https://www.jastt.org/index.php/jasttpath/article/view/65

[10] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, "Concepts of artificial intelligence for computer-assisted drug discovery," vol. 119, no. 18, pp. 10 520–10 594. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.chemrev.8b00728

[11] S. N. H. Bukhari, J. Webber, and A. Mehbodniya, "Decision tree based ensemble machine learning model for the prediction of zika virus t-cell epitopes as potential vaccine candidates," vol. 12, no. 1, p. 7810. [Online]. Available: https://www.nature.com/articles/s41598-022-11731-6

[12] R. Rivera-Lopez, J. Canul-Reich, E. Mezura-Montes, and M. A. Cruz-Chávez, "Induction of decision trees as classification models through metaheuristics," vol. 69, p. 101006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2210650221001681

[13] X. Han, X. Zhu, W. Pedrycz, and Z. Li, "A three-way classification with fuzzy decision trees," vol. 132, p. 109788. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1568494622008377

[14] W. Wegier, M. Koziarski, and M. Wozniak, "Multicriteria classifier ensemble learning for imbalanced data," vol. 10, pp. 16 807–16 818. [Online]. Available: https://ieeexplore.ieee.org/document/9706443/

[15] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm*," vol. 11, no. 2. [Online]. Available: http://thesai.org/Publications/ViewPaper?Volume=11&Issue=2&Code=IJACSA&SerialNo=77

[16] G. Zhang and A. Gionis, "Regularized impurity reduction: accurate decision trees with complexity guarantees," vol. 37, no. 1, pp. 434–475. [Online]. Available: https://link.springer.com/10.1007/s10618-022-00884-7

[17] R. A. Disha and S. Waheed, "Performance analysis of machine learning models for intrusion detection system using gini impurity-based weighted random forest (GIWRF) feature selection technique," vol. 5, no. 1, p. 1. [Online]. Available: https://doi.org/10.1186/s42400-021-00103-8

[18] M. A. Bouke, A. Abdullah, J. Frnda, K. Cengiz, and B. Salah, "BukaGini: A stability-aware gini index feature selection algorithm for robust model performance," vol. 11, pp. 59 386–59 396. [Online]. Available: https://ieeexplore.ieee.org/document/10147829/

[19] Y. Zhang, B. Nie, J. Du, J. Chen, Y. Du, H. Jin, X. Zheng, X. Chen, and Z. Miao, "Feature selection based on neighborhood rough sets and gini index," vol. 9, p. e1711. [Online]. Available: https://peerj.com/articles/cs-1711

[20] G. Zeng, "On impurity functions in decision trees," pp. 1–19. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/03610926.2024.2317359

[21] E. J. Sturm, M. R. Carbone, D. Lu, A. Weichselbaum, and R. M. Konik, "Predicting impurity spectral functions using machine learning," vol. 103, no. 24, p. 245118, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.103.245118

[22] X. Zhao and X. Nie, "Splitting choice and computational complexity analysis of decision trees," vol. 23, no. 10, p. 1241. [Online]. Available: https://www.mdpi.com/1099-4300/23/10/1241

[23] M. Przybyła-Kasperek and S. Aning, "Stop criterion in building decision trees with bagging method for dispersed data," vol. 192, pp. 3560–3569. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050921018688

[24] T.-T. Huynh-Cam, L.-S. Chen, and H. Le, "Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance," vol. 14, no. 11, p. 318. [Online]. Available: https://www.mdpi.com/1999-4893/14/11/318

[25] I. H. Sarker, "Data science and analytics: An overview from data-driven smart computing, decision-making and applications perspective," vol. 2, no. 5, p. 377. [Online]. Available: https://doi.org/10.1007/s42979-021-00765-8

[26] X. Jiang, "Online english teaching course score analysis based on decision tree mining algorithm," vol. 2021, pp. 1–10. [Online]. Available: https://www.hindawi.com/journals/complexity/2021/5577167/

[27] O. Sagi and L. Rokach, "Explainable decision forest: Transforming a decision forest into an interpretable tree," vol. 61, pp. 124–138. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1566253519307869

[28] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," vol. 3, p. 100071. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2772662222000261

[29] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," vol. 192, p. 105400. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169260719308971

[30] S. M. D. A. C. Jayatilake and G. U. Ganegoda, "Involvement of machine learning tools in healthcare decision making," vol. 2021, no. 1, p. 6679512. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/6679512

[31] A. N. Elmachtoub, J. C. N. Liang, and R. McNellis,

"Decision trees for decision-making under the predict-then-optimize framework."

[32] A. Kim and H. Kim, "A new classification tree method with interaction detection capability," vol. 165, p. 107324. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167947321001584

[33] S. Lee, C. Lee, K. G. Mun, and D. Kim, "Decision tree algorithm considering distances between classes," vol. 10, pp. 69 750–69 756. [Online]. Available: https://ieeexplore.ieee.org/document/9810212/

[34] X. Xie, M.-J. Yuan, X. Bai, W. Gao, and Z.-H. Zhou, "On the gini-impurity preservation for privacy random forests." [Online]. Available: https://openreview.net/forum?id=d47iuwOt3j&referrer=%5Bthe%20profile%20of%20Wei%20Gao%5D(%2Fprofile%3Fid%3D~Wei_Gao7)

[35] M. Aria, C. Cuccurullo, and A. Gnasso, "A comparison among interpretative proposals for random forests," vol. 6, p. 100094. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827021000475

[36] J. Velasco-Mata, V. Gonzalez-Castro, E. F. Fernandez, and E. Alegre, "Efficient detection of botnet traffic by features selection and decision trees," vol. 9, pp. 120 567–120 579. [Online]. Available: https://ieeexplore.ieee.org/document/9523853/

[37] R. Dunne, R. Reguant, P. Ramarao-Milne, P. Szul, L. M. Sng, M. Lundberg, N. A. Twine, and D. C. Bauer, "Thresholding gini variable importance with a single-trained random forest: An empirical bayes approach," vol. 21, pp. 4354–4360. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2001037023003082

[38] S. Yu, F. Alesiani, X. Yu, R. Jenssen, and J. C. Principe, "Measuring dependence with matrix-based entropy functional." [Online]. Available: http://arxiv.org/abs/2101.10160

[39] K. Gajowniczek and T. Ząbkowski, "ImbTreeEntropy: An r package for building entropy-based classification trees on imbalanced datasets," vol. 16, p. 100841. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352711021001242

[40] K. Uma and S. A. Alias Balamurugan, "C5.0 decision tree model using tsallis entropy and association function for general and medical dataset," pp. –1–1. [Online]. Available: http://autosoftjournal.net/paperShow.php?paper=100000153

[41] V. Ignatenko, A. Surkov, and S. Koltcov, "Random forests with parametric entropy-based information gains for classification and regression problems," vol. 10, p. e1775. [Online]. Available: https://peerj.com/articles/cs-1775

[42] P. Yu and Y. Deng, "On the scaled rényi entropy and application," vol. 0, no. 0, pp. 1–0, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610926.2024.2301986. [Online]. Available: https://doi.org/10.1080/03610926.2024.2301986

[43] K. Gajowniczek and T. Ząbkowski, "Interactive decision tree learning and decision rule extraction based on the ImbTreeEntropy and ImbTreeAUC packages," vol. 9, no. 7, p. 1107. [Online]. Available: https://www.mdpi.com/2227-9717/9/7/1107

[44] R. P. Bonidia, A. P. Avila Santos, B. L. S. De Almeida, P. F. Stadler, U. Nunes Da Rocha, D. S. Sanches, and A. C. P. L. F. De Carvalho, "Information theory for biological sequence classification: A novel feature extraction technique based on tsallis entropy," vol. 24, no. 10, p. 1398. [Online]. Available: https://www.mdpi.com/1099-4300/24/10/1398

[45] C. Kumar, G. Podestá, K. Kilpatrick, and P. Minnett, "A machine learning approach to estimating the error in satellite sea surface temperature retrievals," vol. 255, p. 112227. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425720306003

[46] A. Hamadani, N. A. Ganai, S. Mudasir, S. Shanaz, S. Alam, and I. Hussain, "Comparison of artificial intelligence algorithms and their ranking for the prediction of genetic merit in sheep," vol. 12, no. 1, p. 18726. [Online]. Available: https://www.nature.com/articles/s41598-022-23499-w

[47] C. J. Nzekwe, S. Kim, and S. A. Mostafa, "Interaction selection and prediction performance in high-dimensional data: A comparative study of statistical and tree-based methods," pp. 259–279. [Online]. Available: https://jds-online.org/doi/10.6339/24-JDS1127

[48] Y. Wu, Y. Ke, Z. Chen, S. Liang, H. Zhao, and H. Hong, "Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping," vol. 187, p. 104396. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0341816219305387

[49] S. Mishra, P. K. Mallick, H. K. Tripathy, A. K. Bhoi, and A. González-Briones, "Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier," vol. 10, no. 22, p. 8137. [Online]. Available: https://www.mdpi.com/2076-3417/10/22/8137

[50] P. Yariyan, S. Janizadeh, T. Van Phong, H. D. Nguyen, R. Costache, H. Van Le, B. T. Pham, B. Pradhan, and J. P. Tiefenbacher, "Improvement of best first decision trees using bagging and dagging ensembles for flood probability mapping," vol. 34, no. 9, pp. 3037–3053. [Online]. Available: https://doi.org/10.1007/s11269-020-02603-7

[51] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model," vol. 2021, pp. 1–11. [Online]. Available: https://www.hindawi.com/journals/complexity/2021/6634811/

[52] V. Aceña, I. Martín De Diego, R. R. Fernández, and J. M. Moguerza, "Minimally overfitted learners: A general framework for ensemble learning," vol. 254, p. 109669. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705122008450

[53] Z. Azam, M. M. Islam, and M. N. Huda, "Comparative analysis of intrusion detection systems and machine learning-based model analysis through decision tree," vol. 11, pp. 80 348–80 391. [Online]. Available: https://ieeexplore.ieee.org/document/10185955/

[54] W. Lian, G. Nie, B. Jia, D. Shi, Q. Fan, and Y. Liang, "An intrusion detection method based on

decision tree-recursive feature elimination in ensemble learning," vol. 2020, pp. 1–15. [Online]. Available: https://www.hindawi.com/journals/mpe/2020/2835023/

[55] X. Luo, X. Wen, M. Zhou, A. Abusorrah, and L. Huang, "Decision-tree-initialized dendritic neuron model for fast and accurate data classification," vol. 33, no. 9, pp. 4173–4183. [Online]. Available: https://ieeexplore.ieee.org/document/9380661/

[56] M. Fernando, F. Cèsar, N. David, and H. José, "Missing the missing values: The ugly duckling of fairness in machine learning," vol. 36, no. 7, pp. 3217–3258. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/int.22415

[57] S. Nijman, A. Leeuwenberg, I. Beekers, I. Verkouter, J. Jacobs, M. Bots, F. Asselbergs, K. Moons, and T. Debray, "Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review," vol. 142, pp. 218–229. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0895435621003759

[58] Z. He, Z. Wu, G. Xu, Y. Liu, and Q. Zou, "Decision tree for sequences," pp. 1–1. [Online]. Available: https://ieeexplore.ieee.org/document/9411677/

[59] A. Simonetto, E. Dall'Anese, S. Paternain, G. Leus, and G. B. Giannakis, "Time-varying convex optimization: Time-structured algorithms and applications," publisher: arXivVersion Number: 1. [Online]. Available: https://arxiv.org/abs/2006.08500

[60] J. Hatwell, M. M. Gaber, and R. M. A. Azad, "CHIRPS: Explaining random forest classification," vol. 53, no. 8, pp. 5747–5788. [Online]. Available: https://link.springer.com/10.1007/s10462-020-09833-6

[61] V. Blanco, A. Japón, and J. Puerto, "Robust optimal classification trees under noisy labels," vol. 16, no. 1, pp. 155–179. [Online]. Available: https://link.springer.com/10.1007/s11634-021-00467-2

[62] S. B. Jabeur, A. Sadaaoui, A. Sghaier, and R. Aloui, "Machine learning models and cost-sensitive decision trees for bond rating prediction," vol. 71, no. 8, pp. 1161–1179. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01605682.2019.1581405

[63] E. García-Martín, N. Lavesson, H. Grahn, E. Casalicchio, and V. Boeva, "Energy-aware very fast decision tree," vol. 11, no. 2, pp. 105–126. [Online]. Available: https://link.springer.com/10.1007/s41060-021-00246-4

[64] R. Lasfar and G. Tóth, "The difference of model robustness assessment using cross-validation and bootstrap methods," vol. 38, no. 6, p. e3530. [Online]. Available: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/cem.3530

[65] N. Syam and R. Kaul, *Random Forest, Bagging, and Boosting of Decision Trees*. Emerald Publishing Limited, pp. 139–182. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/978-1-80043-880-420211006/full/html

[66] V. A. S. Hernández, R. Monroy, M. A. Medina-Pérez, O. Loyola-González, and F. Herrera, "A practical tutorial for decision tree induction: Evaluation measures for candidate splits and opportunities," vol. 54, no. 1, pp. 1–38. [Online]. Available: https://dl.acm.org/doi/10.1145/3429739

[67] S. Bashir, I. U. Khattak, A. Khan, F. H. Khan, A. Gani, and M. Shiraz, "A novel feature selection method for classification of medical data using filters, wrappers, and embedded approaches," vol. 2022, pp. 1–12. [Online]. Available: https://www.hindawi.com/journals/complexity/2022/8190814/

[68] N. Singh and P. Singh, "A hybrid ensemble-filter wrapper feature selection approach for medical data classification," vol. 217, p. 104396. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169743921001647

[69] A. Nugroho, A. Z. Fanani, and G. F. Shidik, "Evaluation of feature selection using wrapper for numeric dataset with random forest algorithm," in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE, pp. 179–183. [Online]. Available: https://ieeexplore.ieee.org/document/9573249/

[70] C. Wu and W. Li, "Enhancing intrusion detection with feature selection and neural network," vol. 36, no. 7, pp. 3087–3105. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/int.22397

[71] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," vol. 7, no. 1, p. 52. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00327-4

[72] Z. Liang, C. Wang, and K. U. J. Khan, "Application and comparison of different ensemble learning machines combining with a novel sampling strategy for shallow landslide susceptibility mapping," vol. 35, no. 6, pp. 1243–1256. [Online]. Available: https://link.springer.com/10.1007/s00477-020-01893-y

[73] D. Arpit, H. Wang, Y. Zhou, and C. Xiong, "Ensemble of averages: Improving model selection and boosting performance in domain generalization," version Number: 4. [Online]. Available: https://arxiv.org/abs/2110.10832

[74] Z. Sun, C. Wang, Y. Zhao, and C. Yan, "Multi-label ECG signal classification based on ensemble classifier," vol. 8, pp. 117 986–117 996. [Online]. Available: https://ieeexplore.ieee.org/document/9125941/

[75] J. Abdollahi and B. Nouri-Moghaddam, "A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation," vol. 5, no. 3, pp. 229–246. [Online]. Available: https://link.springer.com/10.1007/s42044-022-00104-x

[76] T. Zeng, L. Wu, D. Peduto, T. Glade, Y. S. Hayakawa, and K. Yin, "Ensemble learning framework for landslide susceptibility mapping: Different basic classifier and ensemble strategy," vol. 14, no. 6, p. 101645. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1674987123001123

[77] P. Pintelas and I. E. Livieris, "Special issue on ensemble learning and applications," vol. 13, no. 6, p. 140. [Online]. Available: https://www.mdpi.com/1999-4893/13/6/140

[78] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," vol. 10, pp. 99 129–99 149. [Online]. Available: https://ieeexplore.ieee.org/document/9893798/

[79] C. Chen, Y. Tsai, F. Chang, and W. Lin, "Ensemble

feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results," vol. 37, no. 5, p. e12553. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/exsy.12553

[80] G. Tuysuzoglu and D. Birant, "Enhanced bagging (eBagging): A novel approach for ensemble learning," vol. 17, no. 4, pp. 515–528. [Online]. Available: https://iajit.org/PDF/July%202020,%20No.%204/17698.pdf

[81] A. Arabameri, S. Chandra Pal, F. Rezaie, R. Chakrabortty, A. Saha, T. Blaschke, M. Di Napoli, O. Ghorbanzadeh, and P. T. Thi Ngo, "Decision tree based ensemble machine learning approaches for landslide susceptibility mapping," vol. 37, no. 16, pp. 4594–4627. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10106049.2021.1892210

[82] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," *Engineering Reports*, vol. 5, no. 4, p. e12599, Apr. 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/eng2.12599

[83] R. Mao, X. Shi, and Z. Shi, "A Decision Tree Classification Algorithm Based on Two-Term RS-Entropy," *Entropy*, vol. 27, no. 10, p. 1069, Oct. 2025. [Online]. Available: https://www.mdpi.com/1099-4300/27/10/1069

[84] J. He, Z. Li, and L. Yin, "An Efficient and Accurate Random Forest Node-Splitting Algorithm Based on Dynamic Bayesian Methods," *Machine Learning and Knowledge Extraction*, vol. 7, no. 3, p. 70, Jul. 2025. [Online]. Available: https://www.mdpi.com/2504-4990/7/3/70

[85] R. C. Ferreira, K. Dunn-Lopez, S. Moorhead, A. Krupp, B. V. Zuchatti, L. A. C. Carvalho, M. T. D. S. L. Botelho, and E. C. M. Duran, "Decision Trees for Managing Impaired Physical Mobility in Multiple Trauma Patients," *Journal of Advanced Nursing*, p. jan.17010, May 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/jan.17010

[86] A. O. Ciftcioglu, A. Delikanlı, T. Shafighfard, and F. Bagherzadeh, "Machine learning based shear strength prediction in reinforced concrete beams using Levy flight enhanced decision trees," *Scientific Reports*, vol. 15, no. 1, p. 27488, Jul. 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-12359-y

[87] P. Li, F. Xiong, X. Huang, and X. Wen, "Construction and optimization of vending machine decision support system based on improved c4.5 decision tree," vol. 10, no. 3, p. e25024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2405844024010557

[88] T. DeLise, "Era splitting – invariant learning for decision trees." [Online]. Available: http://arxiv.org/abs/2309.14496

[89] S. Marton, S. Lüdtke, C. Bartelt, and H. Stuckenschmidt, "GradTree: Learning axis-aligned decision trees with gradient descent." [Online]. Available: http://arxiv.org/abs/2305.03515

[90] R. C. Arbiv, L. Lovat, A. Rosenfeld, and D. Sarne, "Optimizing decision trees for enhanced human comprehension," in *Artificial Intelligence. ECAI 2023 International Workshops*, S. Nowaczyk, P. Biecek, N. C. Chung, M. Vallati, P. Skruch, J. Jaworek-Korjakowska, S. Parkinson, A. Nikitas, M. Atzmüller, T. Kliegr, U. Schmid, S. Bobek, N. Lavrac, M. Peeters, R. Van Dierendonck, S. Robben, E. Mercier-Laurent, G. Kayakutlu, M. L. Owoc, K. Mason, A. Wahid, P. Bruno, F. Calimeri, F. Cauteruccio, G. Terracina, D. Wolter, J. L. Leidner, M. Kohlhase, and V. Dimitrova, Eds. Springer Nature Switzerland, vol. 1947, pp. 366–381, series Title: Communications in Computer and Information Science. [Online]. Available: https://link.springer.com/10.1007/978-3-031-50396-2_21

[91] D. Loreti and G. Visani, "Parallel approaches for a decision tree-based explainability algorithm," vol. 158, pp. 308–322. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167739X24001857

[92] T. Finke, M. Hein, G. Kasieczka, M. Krämer, A. Mück, P. Prangchaikul, T. Quadfasel, D. Shih, and M. Sommerhalder, "Tree-based algorithms for weakly supervised anomaly detection," vol. 109, no. 3, p. 034033. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.109.034033

[93] J. J. Baena-Rojas, T. M. Mackenzie-Torres, G. Cuesta-Giraldo, and A. Tabares, "A HYBRID MULTI-CRITERIA DECISION-MAKING TECHNIQUE FOR INTERNATIONAL MARKET SELECTION IN SMES," vol. 27, no. 1, pp. 26–45. [Online]. Available: https://pjms.zim.pcz.pl/resources/html/article/details?id=611202

[94] U. J. Ujwal and S. Malik, "A hybrid weight based feature selection algorithm for predicting students' academic advancement by employing data science approaches," vol. 13, no. 5, pp. 1–22. [Online]. Available: https://www.mecs-press.org/ijeme/ijeme-v13-n5/v13n5-1.html

[95] S. Pathan and S. K. Sharma, "Design an optimal decision tree based algorithm to improve model prediction performance," vol. 11, no. 6, pp. 127–133. [Online]. Available: https://ijritcc.org/index.php/ijritcc/article/view/7295

[96] M. K. Subramani and M. G. Muruganantharaj, "EnhancedTree+: A novel approach for improving decision tree classifiers," vol. 11, no. 7, pp. 208–216.

[97] D. Asif, M. Bibi, M. S. Arif, and A. Mukheimer, "Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization," vol. 16, no. 6, p. 308. [Online]. Available: https://www.mdpi.com/1999-4893/16/6/308

[98] S. Cui, A. Sudjianto, A. Zhang, and R. Li, "Enhancing robustness of gradient-boosted decision trees through one-hot encoding and regularization." [Online]. Available: http://arxiv.org/abs/2304.13761

[99] L. L. Custode and G. Iacca, "Evolutionary learning of interpretable decision trees," vol. 11, pp. 6169–6184. [Online]. Available: https://ieeexplore.ieee.org/document/10015004/

[100] M. Riansyah, S. Suwilo, and M. Zarlis, "Improved accuracy in data mining decision tree classification using adaptive boosting (adaboost)," vol. 8, no. 2, pp. 617–622. [Online]. Available: https://jurnal.polgan.ac.id/index.php/sinkron/article/view/12055

[101] H. N. B, A. Jain, and N. Nagaraj. Permutation decision trees. [Online]. Available: https://arxiv.org/abs/2306.02617v3

[102] M. Awad and S. Fraihat, "Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems," vol. 12, no. 5, p. 67. [Online]. Available: https://www.mdpi.com/2224-2708/12/5/67

[103] I. Roshanski, M. Kalech, and L. Rokach, "Automatic feature engineering for learning compact decision trees," *Expert Syst. Appl.*, vol. 229, no. PA, Nov. 2023. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.120470

[104] J. Bodine and D. S. Hochbaum, "A better decision tree: The max-cut decision tree with modified PCA improves accuracy and running time," vol. 3, no. 4, p. 313. [Online]. Available: https://link.springer.com/10.1007/s42979-022-01147-4

[105] M. Z. Horváth, M. N. Müller, M. Fischer, and M. Vechev, "(de-)randomized smoothing for decision stump ensembles," 2022. [Online]. Available: https://arxiv.org/abs/2205.13909

[106] D. Li and S. Chen, "A novel splitting criterion inspired by geometric mean metric learning for decision tree," 2022. [Online]. Available: https://arxiv.org/abs/2204.11011

[107] Y. Manzali, M. El Far, M. Chahhou, and M. Elmohajir, "Enhancing weak nodes in decision tree algorithm using data augmentation," *Cybernetics and Information Technologies*, vol. 22, pp. 50–65, 06 2022.

[108] A. Agarwal, Y. S. Tan, O. Ronen, C. Singh, and B. Yu, "Hierarchical shrinkage: improving the accuracy and interpretability of tree-based methods," *CoRR*, vol. abs/2202.00858, 2022. [Online]. Available: https://arxiv.org/abs/2202.00858